DOCUMENT RESUME

ED 271 648 CG 019 187

AUTHOR Kirkpatrick, J. Stephen
TITLE The Magic of Structure.
INSTITUTION Planned Parenthood Federation of America, Inc., New York, NY. Education Dept.
REPORT NO ISBN-0-934586-06-3
PUB DATE 81
NOTE 257p.
AVAILABLE FROM Education Dept., Planned Parenthood Federation of America, Inc., 810 Seventh Avenue, New York, NY 10019.
PUB TYPE Guides - Non-Classroom Use (055)

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS *Family Life Education; *Family Planning; *Program Development; *Program Evaluation; *Reference Materials; *Sex Education
IDENTIFIERS *Cyclical Evaluation Model

ABSTRACT

Part I of this two-part manual contains three sections on procedures and processes for planning and evaluating family planning education programs. Section 1 contains a working definition of evaluation and a description of its relationship to research, a presentation of the cyclical evaluation model upon which this manual is based, reasons why evaluation and planning are important, and an explanation of how this manual is structured. Section 2 explains how to devise a plan of action, write lists of priorities and goals, write behavioral objectives, evaluate objectives, and set standards in objectives. Section 3 presents the eight-step process of the cyclical evaluation model, explains the needs assessment process, examines the use of behavioral objectives, and discusses various aspects of program evaluation. The seven sections in Part II of this manual provide sources of reference materials and practical suggestions on the techniques and methodologies needed to implement the process outlined in Part I. The seven sections are: (1) locating and using other people's data; (2) how to write questionnaires; (3) writing tests; (4) surveys; (5) experimental research; (6) analyzing data; and (7) a thorough example. This document concludes with a selected bibliography, an index, and a glossary of terms. It is noted that, although the manual specifically refers to family planning education, the materials presented can be used in the fields of sexuality education and health education in general. (NB)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

The Magic of Structure

ED271648

A Manual for Planning and Evaluating Family Planning Education Programs

J. STEPHEN KIRKPATRICK, Ph.D.
Program Development Specialist
DEPARTMENT OF EDUCATION

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Gloria A. Roberts

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

CG 019187

Planned Parenthood Federation of America
810 Seventh Avenue
New York, New York 10019

1981

Library of Congress Cataloging in Publication Data

Kirkpatrick, J. Stephen
The magic of structure.

Biblography
Includes index.
1. Birth control--Study and teaching--United States. 2. Birth control--Study and teaching--United States--Evaluation. I. Title.
HQ763.6.U5K57 613.9'4'07 81-1824
ISBN 0-934586-06-3 AACR2

KDM Library Classification #8.1 K5 m

Copyright © 1981
Planned Parenthood Federation of America, Inc.

All rights reserved.

PREFACE AND ACKNOWLEDGEMENTS

Part I, the first three sections of the ten in this manual, contains the backbone of the procedures and processes which I suggest for planning and evaluating family planning education programs. Thus, while the manual may appear somewhat intimidating in terms of size, the rest of the manual is intended as a source of practical suggestions on the various techniques and methodologies needed to implement the process outlined in the beginning sections. I included the seven sections in Part II as a source of reference material.

Throughout this manual I have tried to assume as little prior knowledge on the part of the reader as space would permit. Some family planning educators may have an extensive familiarity and vast experience with one or more of the topics I present here. However, the beginner has been kept in mind throughout the process of writing this manual.

Family planning educators come from a variety of educational and professional backgrounds, bringing with them widely varying skills, talents, and experiences. Thus, to facilitate use of this manual, each section begins with a description of its contents. Seven of the ten sections end with a self-test, to aid in reviewing material you may already be familiar with, or to check your learning process.

References in the text follow the author-year method of citation, rather than footnotes, with the full citation listed in the bibliography.

While I have used the term "family planning education" throughout the manual, the system which I presented in Part I and the material in Part II apply just as well to the whole field of sexuality education and training. In fact, the system will apply quite well to the entire field of health education as attempted by agencies and institutions.

Each subdivision of the sections in the manual has a numerical label. For example, the major subdivisions of section one are numbered 1.1, 1.2, 1.3, and 1.4. Sometimes a subdivision is further divided, as in 9.2.1, 9.2.2, and so forth. That system facilitates cross-referencing. To enable you to find cross-references, each page is labeled with the subdivisions. The index contains page number citations.

Professionals who use this manual will embed its suggestions into the different management styles found within their agencies and within varying traditions of getting things done. Those styles and traditions will influence how the various suggestions and systems in the manual are put into use. I have tried to emphasize the need for structure and system, not to prescribe any particular style of management or suggest any hard and fast rules for separating management and administrative decisions from policy-making. I have emphasized system, structure, and communication, which are essential no matter what style of management an agency chooses to use, and no matter how the various roles and functions are divided among educators, administrators, and policy-makers.

None of the examples given in the text should be construed as prescriptions of either method or philosophy; they are intended to illustrate

concepts or clarify principles. And, although I have tried to offer realistic and potentially helpful examples, they reflect my own experiences and prejudices.

Because the field of planning, evaluation, statistics, and research includes so much information and so many concepts, I encourage you to rely upon consultants and technical assistance from experts whenever you need them. Seek technical assistance from others frequently. If you work for the Planned Parenthood Federation of America (PPFA) or one of its affiliates, or if you serve on the board of directors for the Federation or one of its affiliates, do not hesitate to call the PPFA Education Department for technical assistance.

* * * * * *

Jane Watkins from the staff of PPFA and Lynne Cooper of Planned Parenthood of Santa Cruz County, Calif., offered invaluable comments on the first three sections of the manual. Their feedback was cogent and indispensable to the writing and editing process.

John Stahle from the Communications Division of PPFA designed the cover and did the artwork for the figures.

I also want to thank the following educators and their affiliates for assisting me by generously agreeing to serve as field-test sites for the final draft of **The Magic of Structure:**

Mary Berlin, Planned Parenthood Association of Nashville, Tenn.;
John Cunico, Planned Parenthood of Minnesota, St. Paul, Minn.;
Brenda W. Hansen, Planned Parenthood of Southern Arizona, Tucson, Ariz.;
Kathy Kowalski, Planned Parenthood of Alameda-San Francisco, S.F., Calif.;
Jan Loreen, Planned Parenthood of Seattle/King Co., Wash.;
Marylyn Lyman and Linda Oken, Planned Parenthood of Syracuse, N.Y.;
Karla Myers, Planned Parenthood of Central Missouri, Columbia, Mo.:
Sallye Puryear, Planned Parenthood of Southeastern Pennsylvania, Philadelphia, Penn.;
Meg Smillie, Reproductive Health Care Centers of South Central Michigan, an Affiliate of the Planned Parenthood Federation of America, Kalamazoo, Mich.;
Sarah Spengler, Planned Parenthood of Metropolitan Washington, D.C.;
Sara Traphagen, Planned Parenthood of Humboldt Co., Eureka, Calif.;
Esther Walter, Planned Parenthood of Mid-Iowa, Des Moines, Iowa;

and to thank the co-workers and board members from those affiliates who also reviewed the manual. Identifying them all would be impossible, yet they deserve as much gratitude as those I have named.

Thanks also to Carol Cassell, Timothy P. Lannan, Carol A. Pratt, and Marilyn Iris Auerbach for all of their assistance.

JSK
New York, N.Y.
December, 1980

THE MAGIC OF STRUCTURE

TABLE OF CONTENTS

Part One: THE STRUCTURE OF PLANNING AND EVALUATION

section one: WHAT AND WHY 1

1.1 What 3
1.2 Evaluation models 4
1.3 Why do evaluations 6
1.4 The remainder of this manual 8

section two: RESOURCE MANAGEMENT:
SETTING PRIORITIES, AND WRITING GOALS AND OBJECTIVES . . . 9

2.1 Who. 10
2.2 Writing behavioral objectives. 15
2.2.1 The easy objectives 15
2.2.2 "Outcome" objectives. 17
2.2.3 Pitfalls. 19
2.2.4 Some harder ones. 20
2.3 Relating objectives to reality 24

section three: THE PLANNING AND EVALUATION PROCESS. 35

3.1 What to evaluate 36
3.2 The assessment phase 39
3.3 The planning phase 42
3.3.1 Step 2. 42
3.3.2 Step 3. 42
3.3.3 Step 4. 45
3.3.4 Step 5. 47
3.4 Implementation (Step 6). 52
3.5 Evaluation 52
3.5.1 Step 7. 52
3.5.2 Step 8. 53
3.6 Consultants. 55

Part Two: METHODS AND TECHNIQUES

section four: LOCATING AND USING OTHER PEOPLE'S DATA 61
4.1 Sources of data. 63
4.2 Using these data 65
4.3 The question of impact 69

section five: HOW TO WRITE QUESTIONNAIRES 72

5.1 Guidelines for constructing questionnaires 73
5.2 Item types 80
5.3 Pretesting 88
5.4 Interpreting questionnaire results 90

section six: WRITING TESTS. 97

6.1 Evaluation vs. measurement 98
6.2 Validity and reliability 100
6.2.1. Valitity 100
6.2.2. Reliability 103
6.3 Construction of a test 108
6.3.1 Tables of specifications. 109
6.3.2 Determining the number and types of items 111
6.3.3 Writing and choosing items. 113
6.3.4 Preparing the test papers 114
6.4 Item types 115
6.4.1 Multiple-choice 115
6.4.2 True-false. 119
6.4.3 Dichotomous-choice. 120
6.4.4 Cloze items 121
6.4.5 Matching items. 121
6.4.6 Hierarchical recognition. 122
6.4.7 Interpretive task 123
6.5 Scoring and interpreting 123
6.5.1 Mechanics of scoring. 123
6.5.2 Item analysis 124
6.5.3 Correction for guessing 126
6.5.4 Interpreting test scores. 128
6.6 Using tests constructed by others. 128

section seven: SURVEYS. 134

7.1 Types of surveys 135
7.2 Steps in conducting a survey 138
7.3 Sampling 142
7.4 Other people's questionnaires. 147

section eight: EXPERIMENTAL RESEARCH 151

8.1 Sample research problems and design 153
8.2 Threats to experimental validity. 156
8.2.1 Threats to internal validity 157
8.2.2 Threats to external validity 160
8.3 Analysis of the results 164

section nine: ANALYZING DATA 170

9.1 Levels of measurement 172
9.2 Descriptive statistics. 175
9.2.1 Presentation of raw data 175
9.2.2 Percentages. 179
9.2.3 Measures of central tendency 180
9.2.4 Measures of dispersion 185
9.3 Inferential statistics. 191
9.3.1 Parametric vs. nonparametric statistics. 195
9.3.2 Measures of association (correlation). 197
9.3.3 Comparisons of groups. 206
9.3.4 The case of one or multiple groups 207
9.3.5 Multivariate statistics. 208

section ten: A THOROUGH EXAMPLE. 214

10.1 Background information. 215
10.1.1 Setting priorities. 216
10.1.2 Goals 216
10.2 Planning the education program. 217
10.3 Implementation. 220
10.4 Evaluation. 221
10.5 Sample forms. 224

section eleven: SELECTED BIBLIOGRAPHY. 229

section twelve: INDEX. 234
GLOSSARY 237

section one:
WHAT AND WHY

This section contains:

1. A working definition of evaluation, along with a description of its relationship to research.

2. A graphic presentation and description of the cyclical evaluation model, the system of planning and evaluation upon which this manual is based.

3. Seven reasons why evaluating and planning are important, if not indispensable, to family planning educators, their supervisors, and their boards of directors.

4. An explanation of how this manual is structured, with goals or objectives for each section on the interleaves (like this one) and self-tests to assist you in reviewing your earlier learning and your work with this manual.

If you are already familiar with evaluation and the importance of evaluation and planning in the educational process, you may want to skip to **section two.**

section one: WHAT AND WHY

"If I had five minutes to solve a problem, I'd spend the first three minutes planning what to do with the last two."

-- Thomas Edison

Evaluation has an atrocious reputation. People view it variously as a threat, a waste (of time, of money, of paper, of energy), a pain in various parts of the anatomy, an exercise in futility, a way to avoid decisions, the forerunner of unwanted change, and an onerous task. The word conjures up recollections or fantasies of awkward interviews with the boss, sleepless nights, and damp palms. To evaluate, it seems, is the sentence issued by a group of demoniacally possessed bureaucrats sitting around a simmering cauldron of paperwork headaches. And when the prefix "re-" appears in front of the word, as in re-evaluate, fear clouds reason, doom seems imminent, and suicide becomes attractive.

Consequently, the first task here must be to sell you the notion that evaluation is none of those things, unless someone abuses it or you. Instead, evaluation can provide the exact opposite of those things: security, economy, expedience, efficiency, information, and restful satisfaction.

Perhaps you're thinking, "Look, I'm busy. I have a thousand things to do, one hundred of them by five o'clock--yesterday! Give me one good reason why I should read even the rest of this page, let alone the rest of this manual."

There are at least seven good reasons why this manual can prove useful, even invaluable, to y[illegible] However, before stating those, it's necessary to define evaluation.

If you will take the time to read the rest of this section, about fifteen minutes or so, I will present seven reasons for evaluating, as well as some goods news about how to use the rest of this manual without having to read all of it at once--perhaps ever. In fact, Part I (this section and the next two) contains all of the structure for planning and evaluation. Part II (sections four through ten) covers methods and techniques you will use only as you need to refer to them.

1.1 WHAT

Research is any defensible, systematic process of collecting data which finds an answer to a specific question, an answer which, ideally, furthers human knowledge. The research process involves applying various kinds of skills and tools to a situation in order to collect data. As used here, "data" means any kind of facts or information, whether numbers, statistics, ideas, references, objects, photographs, charts, graphs, etc. Process, however, is the essense of the definition. Research is neither an isolated event nor a randomly occurring series of disconnected events, but a well-planned series of stages, steps, or events logically connected to, and leading toward, an answer.

Evaluation is research or some other data-collection process designed to enable informed judgments, to assign values to a performance or product--to E-VALU-ATE. While it may further human knowledge, evaluation leads primarily to conclusions about a program, activity, or event in terms of its value: good or bad, effective or ineffective, efficient or inefficient; worthy or unworthy of funding.

Evaluation and research differ also with respect to criteria for outcome decisions. A research project either does or does not answer a specific question. A project designed to find the location of Atlantis or a cure for herpes genitalis either does those things or it does not. The value assigned to the data at the conclusion of evaluation research, on the other hand, is usually much more subjective; hence the sweaty palms and sleepless nights. But that does not have to be so. Professional educators can, and perhaps must, define observable criteria which determine whether a program is efficient or not, effective or not, worthy or not.

Those criteria are called **objectives,** a term which also has a bad reputation. But writing objectives and using them as evaluative criteria is neither good nor bad intrinsically. Written objectives can serve as tools or shackles, depending upon how people use the objectives. (See **section two** and, especially, 2.2.1.)

Exactly what the objectives should include, and how they should be stated, depends upon the purpose of the evaluation. Evaluations can determine which manner of doing things is more effective or efficient, how to allocate scarce resources to meet the most crying needs, and whether to fund a program or continue funding it. Unfortunately, some evaluations

serve only to satisfy grant requirements, to delay or avoid difficult decisions, or to collect evidence against a program, agency, or individual.

Every evaluation process includes three roles: the person(s) whose product or performance comes under examination; the evaluator, who designs methods and strategies and collects data; and a decision-maker who assigns a value to the data. One individual may perform two or all three of those roles. At other times, different people, groups, or committees may fulfill one or more of the three roles. The evaluator, the decision-maker(s), or both may work for the same agency as the performers or they may come from government agencies or consulting firms.

A clear and explicit understanding of who will perform which functions and for what purposes is critical from the outset--from the time when the first plans are made to execute a program or produce a product. This manual contains methods, information, and skills useful to all three kinds of people.

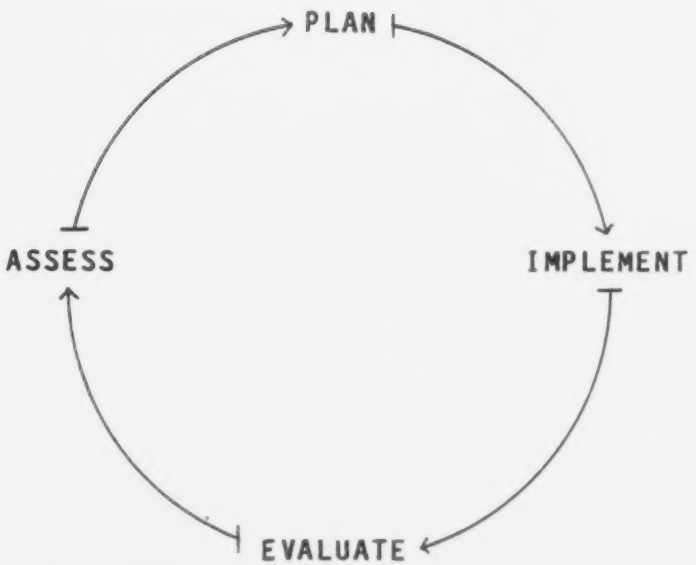
1.2 EVALUATION MODELS

The literature on evaluation includes at least a half-dozen models and philosophies on conducting evaluations. The terms **formative** and **summative** appear in virtually all evaluation textbooks. You do a formative evaluation when you gather information about what a given agency or population needs (conduct a **needs assessment)** and then draw conclusions aimed at program development or improvement. A summative evaluation leads to conclusions, after the fact, about the impact or effectiveness of a program or product.

Family planning education requires frequent repetition of the same or a similar program. Therefore, this manual advocates and describes a **cyclical evaluation model,** represented by **Figure 1.** Ideally, you enter the cycle at the assessment phase, gathering data about the needs of the population, community, or individual at whom a program will be aimed. Based upon that data you formulate a plan to meet those needs, in terms of written objectives. Simultaneously with planning for the educational program, you formulate evaluative strategies and tools. Then, after implementing your educational and evaluative plans, you draw evaluative conclusions--which become a part of your next needs-assessment and planning processes.

Every family planning educator probably GOES THROUGH ALL FOUR PHASES OF THE CYCLICAL EVALUATION MODEL EACH TIME they perform, even if only haphaz-

FIGURE 1. CYCLICAL EVALUATION MODEL



Ideally, one enters the cycle at the **assessment** phase and uses the assessment data to formulate a plan for **implementation** of the program based upon written behavioral objectives. Then one evaluates the implementation of the program. This allows the use of the **evaluation** to provide additional data for the next assessment process and the cycle repeats itself. Sometimes family planning educators must enter the cycle at another phase. In that case, the cycle can still continue; it will simply provide a less complete set of findings until one has followed the cycle back to the assessment process.

A formative evaluation model would include the assessment and planning phases, terminating with recommendations for implementation.

A summative evaluation model would begin at either the implementation phase and continue with evaluation or would consist solely of the evaluation process for a program already completed. Obviously, a summative evaluation can proceed more easily and productively if the evaluator is able to participate in the implementation process, if only to help formulate objectives for use when the time to evaluate arrives.

ardly and informally. By using the tools and skills presented here you can perform these steps more systematically, reduce the subjectivity and anxiety associated with evaluation, and gain the benefits described below.

1.3 WHY DO EVALUATIONS?

This manual can provide seven kinds of valuable benefits: improved accountability, more effective resource management, decreased burn-out, better planning, stronger defense against attacks, increased fund-raising, and personnel evaluations that are more specific and fair.

Accountability. The US DH & HS (Department of Health and Health Services); state, county, and local governments; health systems agencies; charities; philanthropists; and donors have begun to ask for more and more justifications for funding as money gets more scarce. Although we family planning educators may believe, even know, that our work is effective, valuable, and indispensable, we must begin to prove those claims.

And our peers, administrators, boards of directors, and clients will insist upon proof as well--and they may be more strict and insistent than anyone else. We must also begin to arm our allies--those who support our efforts in academic and political arenas--with hard data and convincing evidence.

Resource management. Edison's quote was chosen to begin this section because we face a seemingly insurmountable task: too many people in need, too few of us to serve them all, and too few resources. The cyclical evaluation model will improve the efficiency and effectiveness of what we do so that we can make maximal use of our time and other resources. To achieve this end, you have to accept the idea that **evaluation is a part of the educational process,** and neither an afterthought, a nicety, nor a simple professional obligation. Evaluation and planning are not separate processes, but inseparable parts of one whole.

Burn-out. Depression, anxiety, restlessness, high blood pressure, insomnia, absenteeism, and illness collectively indicate burn-out and that a professional needs some time to "recharge the batteries." Co-workers notice moodiness; routine chores seem like drudgery; nothing seems to matter very much; no effort seems to produce satisfaction. Burnt-out educators often believe that their administrators and boards of directors fail to appreciate or support their work.

Specific causes of burn-out vary depending on setting, characters, and the operation of Murphy's Laws. Projects, problems, and pressures arise daily for family planning educators as surely as earthquakes--and just as unpredictably in terms of when and where they will strike and how much damage they will do.

However, part of the problem is internal; we have a strong tendency to say yes when we would like to say--or should say-- NO!

The cyclical evaluation model can provide you with a clearer set of expectations for yourself and for your administrators and board members. It can give you a sense of satisfaction when you have done a good job. The planning and evaluation process can also help us work with administrators and policy-makers to set priorities and strike a reasonable balance among what has to be done, what most needs to be done, and what we would like to be able to do.

Program planning. Whether we perform a single speaking engagement, conduct a three-day workshop, or design a system for educating and orienting new patients, there are several questions which we should ask ourselves: (1) Did the program meet the needs of the participants? (2) Did the program do what it was supposed to do, have the effect is was designed to have? (3) If we do it again, how can we improve it? (4) How can we prove that the approaches and materials we use are working? (5) Should we discard our current approach and try something different?

Only systematic and conscientious application of the planning and evaluation process can reduce or eliminate the guesswork in answering these questions. More importantly, planning and evaluation can help us share our successes and failures among ourselves, prevent repetition of the same mistakes, and help our profession advance.

Defense against attack. Virtually anyone associated with family planning has a familiarity with the opposition's attacks and tactics. They use any opportunity and forum to harass, embarrass, and persecute us. Their "evidence" consists more often of piety, dogma, and hearsay than facts, data, or pragmatism. Their devotion is absolute; their organization sophisticated; their approach relentless. They have persisted, in one form or another, since the inception of the family planning movement, and they will continue indefinitely.

Education offers them a fertile ground for attacking family planning--until we gather more documented evidence--hard data--to refute and ridicule their claims.

More importantly, evaluative evidence can provide a solid basis from which we can take the initiative and announce our successes to the public. Survey after survey shows that family planning has a good and solid reputation among the public. Our responsibility dictates--and our survival requires--that we be prepared.

Fund-raising. Any experienced fund-raiser will tell you that success speaks convincingly to donors. Family planning educators who can **prove** the success and effectiveness of their programs can arm fund-raisers with very convincing proposals.

Moreover, family planning educators can become fund-raisers' partners, and even fund-raisers themselves, with good evaluative data.

Personnel evaluations. The systematic approach to planning and evaluation presented here can help you in specifying what you expect of people whom you supervise and lets them know how they will be evaluated. Not only will this improve morale and add productive substance to personnel evaluation interviews, but also, in those unfortunate cases when an agency must consider terminating an employee, it will provide substantial evidence of fairness and due process. For a description of how to approach some of the more subjective aspects of personnel performance, see 2.2.4.

1.4 THE REMAINDER OF THIS MANUAL

The remainder of **The Magic of Structure** offers you both an opportunity to learn the skills needed to plan and evaluate your educational programs and a reference source for writing tests and questionnaires, making use of available data, conducting surveys and experiments, and analyzing your data.

Each section begins with a description of its contents, so that you can more easily determine how much of the material you already know. Most of the sections end with a self-test, as a review or an aid in identifying areas you need to study more closely. Please think of the self-test as a choice, not a chore. I hope you will use the manual by reading the first three sections and then referring to the others as you need to use the information in them.

section two:
RESOURCE MANAGEMENT:
SETTING PRIORITIES, AND WRITING GOALS AND OBJECTIVES

After reading this section and learning the material in it you will be able to:

1. Define the meaning of "priorities," "goals," and "objectives."

2. Devise a specific plan of action to prepare priorities, goals, and objectives for your educational program, including the roles and responsibilities of educators, administrators, and board members.

3. Write lists of priorities, goals, and objectives.

4. Write behavioral objectives in various areas of educational programming, including both the factual and attitudinal domains.

5. Evaluate objectives to ascertain that they are, in fact, behavioral objectives.

6. Set standards in your objectives which are meaningful and defensible.

If you believe that you already know this material and can perform the tasks outlined by the above objectives, then you may want to skip this section and go on to **section three.** If you are not sure, or if it has been a while, you may want to turn to the self-test in the back of this section to review what you have learned.

NOTA BENE: IT IS ESSENTIAL THAT YOU KNOW, UNDERSTAND, AND BE FLUENTLY FAMILIAR WITH **ALL** OF THE MATERIAL IN THIS SECTION IN ORDER FOR YOU TO BE ABLE TO UNDERSTAND AND USE THE MATERIALS IN THE SECTIONS TO FOLLOW. The remaining sections make frequent reference to the material contained in section two, and assume that you are fluently familiar with it.

HOWEVER, before you actually put any of the material in section two into practice, you need to read section three: The Planning and Evaluation Process. That section bridges the dangerous ravine between writing objectives and ensuring that you can measure their accomplishment efficiently but acceptably.

section two: RESOURCE MANAGEMENT: SETTING PRIORITIES, AND WRITING GOALS AND OBJECTIVES

> "...If you can't tell one when you see one, you may wind up without a leg to stand on."
>
> --R.F. Mager (1972, p. vi)

This section covers those activities related to evaluation which are most crucial and yet most often forgotten--setting priorities, and writing goals and objectives. Family planning educators frequently find it impossible to ascertain what all of their priorities, goals, and objectives will be for the coming year, or even for some shorter time period. However, in the absence of some attempt at making explicit plans, we have a tendency to expect too much of ourselves and to allow others to expect too much of us.

Once again, I must begin by selling you on the idea that learning how to set priorities and to write goals and objectives is more of a help than a bother. Professionals working in various areas, including business, industry, and human services, have become the victims of objectives and objective writing. Such unpleasant experiences are the fault of the people who abuse objectives; they are not caused by the process of setting objectives and using them to guide daily activities and short- and long-range planning.

The process which I will outline in this section will also prove invaluable because it requires the administration, board of directors, and the educators to work together to communicate specifically about the educational program. Keeping everyone informed about your educational program may be the most valuable outcome of the entire process!

2.1 WHO

Various individuals in any family planning agency structure can and should know what educators are doing and why. I advocate that each of several groups plays some role in the process of setting priorities, goals, and objectives.

The process described here is not intended to be a prescription. Each individual family planning agency must make its own choices regarding the various roles and functions of each of these groups and individuals, based not only upon the size of the agency and its administrative and legal structure, but also upon its traditions and management styles.

Notwithstanding the inevitable variations, wisdom dictates at least some involvement by each of the groups listed below.

While the chronological order listed below is based upon careful thought, it may be necessary for several of the groups to get involved in the process at any one time, or for their input to be included at varying times in the process. The flow chart in **Figure 2** (on the next page) shows how this ideal process would work.

Education staff. The education staff members should set aside a minimum of one-half working day and go to a place where they can work undisturbed, probably to a place away from their agency. They should begin by deciding what their priorities will be for the coming program period (month, quarter, year). They might begin by simply listing things which may be priorities without attempting to place them in any order.

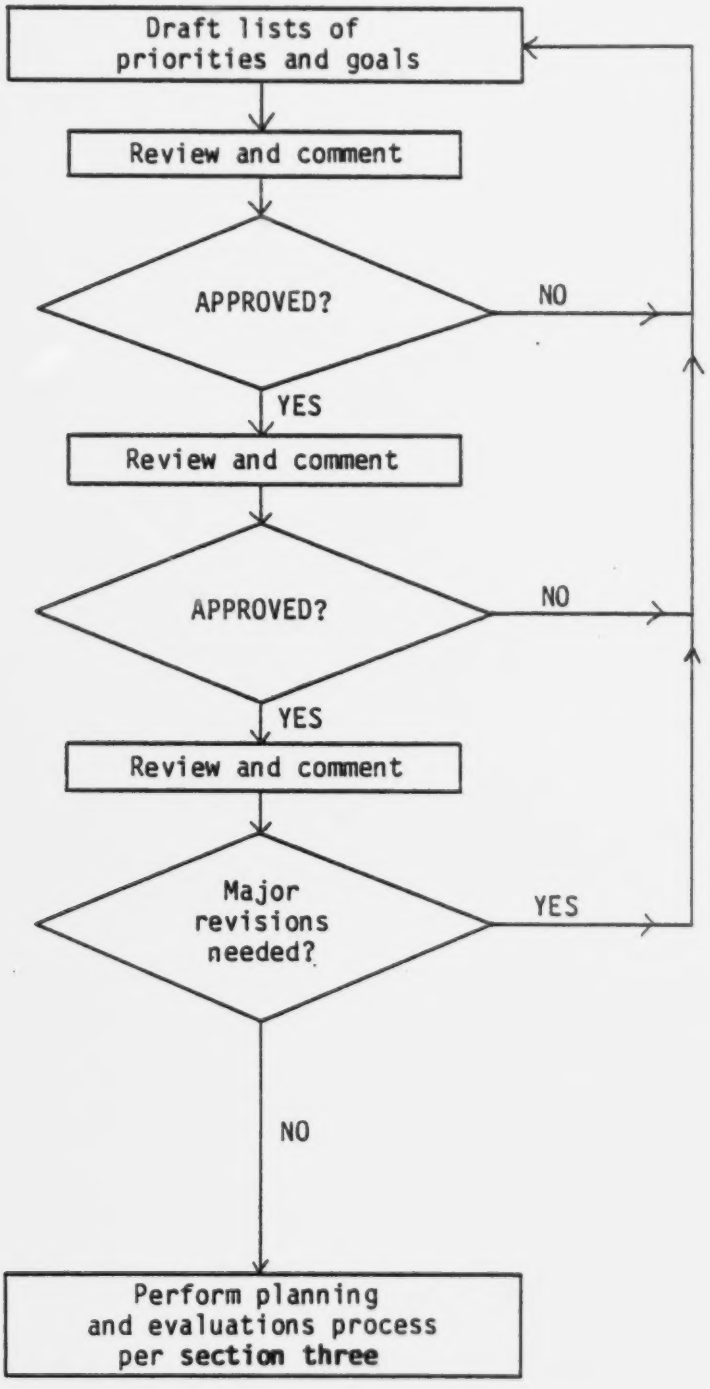
A list of **priorities** is nothing more than a list of either groups or concerns which will occupy the time and focus the efforts of the educators for the next program period. The following lists include examples of groups and concerns, or problems, which might be included in a list of priorities:

GROUPS	CONCERNS
children	networks/community linkages
teenagers	unplanned pregnancies
pre-teens	pregnancies among single women
single mothers, fathers	venereal disease
low-income populations	repeat unplanned pregnancies
minority groups	rape
males	incest
immigrants	family life education
migrant workers	child abuse
post-abortion patients	spouse abuse
pregnancy-test patients	DES exposure
transients	infertility
	public affairs

Some staffs even write lists of priorities in terms of both populations and types of concerns or problems which those populations have.

The listing of priorities is an essential initial step for several reasons. First, and most importantly, it gives everyone involved a clear view of how the family planning educators will spend time, money, and other resources. Second, the process requires at least some informal assessment of the community's needs (see 3.2). Third, it makes a statement about the agency's mission and its hopes for contributing to the community.

Figure 2. Flow Chart: Setting Priorities, Writing Goals and Objectives



Setting priorities should begin with the agency's mission statement, the document which sets forth the agency's reason for existing. Educators should also consult their job description(s). Family planning educators have widely varying responsibilities, ranging from patient recruitment and education to public affairs, public relations, marketing, media relations, and so forth. The priorities and goals should account for each area of responsibility.

Next, the educators should begin to write goals which they hope to reach for each priority. A goal is nothing more than a global statement of a desired outcome. For example, if teenagers are the number-one priority group then the following goals may reflect those hopes:

Through our education efforts we will:

1. increase use of non-prescription contraceptives by sexually active teens;

2. reduce the number of repeat pregnancies among teenagers still attending school;

3. increase the number of new sexually active teenage patients coming to our clinics for birth control information and contraceptive methods;

4. increase contraceptive use among teenage patients who come to our clinics seeking pregnancy-testing services.

These goals present a clear statement of the educators' intentions, which will tend to get more concrete reviews and comments from others than a set of nebulous hopes. I recommend that educators share their priorities and goals with the other groups listed below **before they write behavioral objectives.**

Executive director. The education director should set aside a period of at least an hour or two for a discussion with the executive director (chief administrative officer, president, director...) about the priorities and goals which the education staff has generated. Both the educator and the administrator should openly discuss the relationship between the education program and the agency's other efforts.

When the education director and the executive director have agreed upon both the content and relative order of importance in both the lists of priorities and goals, they should agree upon which, if not both of them, will convey the contents of their draft lists to the next stage.

Board of directors. At this point the education director, the executive director, or both should set aside a time to discuss the lists of priorities and goals with the chairperson of the education committee or the entire committee. I believe strongly in early and consistent involvement of the entire education committee beginning with a review of the priorities and goals. Some agencies choose not to involve the entire committee until the specific objectives are written. In any case, the education committee should have direct involvement in the final lists of objectives, (1) so that they can make the appropriate policy decisions and (2) so that they can work with other committees and the entire board in the planning and budgeting process.

Community. At this point the agency is prepared to discuss its plans with the rest of the community. Some agencies merely mail a list of their priorities and goals to other agencies involved in family planning, health education, family life education, or human services. Other agencies invite representatives from the community to attend an open house or similar forum for discussing the priorities and goals. The methodology is not important; the essential step is to get community input. Frequently, family planning educators discover during this process that other agencies have plans for similar kinds of programs. This enables agencies and programs working in the same community to combine their efforts or, minimally, to avoid duplication of effort.

In the most ideal circumstances we would include input from those populations whom we intend to serve. I believe it is essential for us to remember that our reason for being is to meet the needs of the various groups in our communities. We cannot be entirely certain of those needs unless we communicate directly with those whom we wish to serve.

At the very least, any other individual or agency which offers family planning should have the opportunity to review and comment. Otherwise, some duplication of effort is almost inevitable and, what can be much worse, others may view your efforts as an invasion of their turf.

This will take some period of time varying in length from thirty days to three months. Consequently you need to begin this process of planning at least three months before the beginning of the program period for which you are writing priorities, goals, and objectives.

Once each of these groups--the agency's administration, the agency's board of directors, and groups from the community--have had an opportunity to review and comment upon the educators' priorities and goals, the process of writing objectives can begin.

2.2 WRITING BEHAVIORAL OBJECTIVES

How nice it would be if there were some other way for us family planning educators to define what we do, how we do it, and to prove when we have done it. Unfortunately, writing objectives is the only explicitly defensible and concretely definable method currently available.

A behavioral objective is an explicit statement which contains several important elements: who will do what, how, when, and where, and how one can tell when all of this has been done.

The crucial difference between an objective and a goal is that a trained observer can perceive whether an objective has been met or not through the senses; s/he can touch, taste, hear, see, or smell all that is necessary to evaluate success or failure in the attempted accomplishment of an objective. The evaluation of how well a goal has been accomplished requires some subjective judgment. This kind of standard may seem impossible to meet in some circumstances, especially when educators deal with affective areas--those areas having to do with attitudes, feelings, self-concepts, or observable behavior seemingly too complex to adequately describe in written language. It may even seem difficult to be able to quantify objective outcomes in such areas as "recruiting teenage patients," or "increasing adolescent use of contraceptives."

However, in the overwhelming majority of cases, such objectives can be written without your becoming a slave to a set of objectives or to the process of writing them.

The remainder of this section includes (1) an introduction to writing behavioral objectives; (2) a description of writing "outcome" objectives, statements about what a program should give its participants; (3) some important concepts regarding relating objectives to reality and setting the standards which objectives imply.

2.2.1 THE EASY OBJECTIVES

Writing a set of behavioral objectives for a list of priorities and goals may seem like an onerous, if not impossible, task. There is no doubt that writing behavioral objectives can be hard work from the point of view

that it requires concentration. Dividing the job among those who will be responsible for accomplishing each set of objectives lightens the load and helps each educator understand his or her tasks.

Some people write objectives in three component parts: a rationale, the objective, and a measurement strategy. For example, suppose all of the various persons associated with an agency have agreed on the following:

Goal: We want to ensure the quality of our new-patient education program.

Then, you might write the following objective related to the goal:

Rationale: Since the Directors' Council has decided that new-patient classes will be required for all new patients, a method of evaluation for new-patient classes is desirable and necessary.

Objective: During the months of June, July, and August, 1980, we will distribute questionnaires to all participants in our new-patient classes held during those months for the purpose of determining how much the patients are learning and how the patients perceive the instructional methods used.

Measurement: The Education Director will compile the data from the questionnaires, write a report which includes suggestions for improving the new-patient education system, and submit it to the Executive Director and the Directors' Council by September 30, 1980.

Formal inclusion of the three elements in the above example is not always necessary. You might choose to omit the rationale, assuming that it is obvious to all concerned. Or, you might choose to state the objective and how it will be measured in one sentence. For example, one might have written the above objective in the following statement:

Objective: During the months of June, July, and August, 1980, the education staff and volunteers will distribute questionnaires to all participants in new-patient classes to determine how much the patients are learning and how the patients perceive the instructional methods used, and they will submit a report on their findings to the Executive Director by September 30, 1980.

An objective does not have to be such a lengthy statement as the example above. For example, suppose one of your education goals for the coming program period is to increase the number of teenage patients seeking contraceptives in your family planning clinic. Then, you might have the following list of objectives aimed at accomplishing that goal:

1. We will increase our number of speaking contacts with teenagers from the 8,500 we had in 1979 to 10,000 during 1980.

2. We will secure air-time for a total of 360 public service announcements during 1980 on local stations KROC, KINT, KCUF, KFLT, and KRYS.

3. We will create, print, and distibute 500 posters during 1980 aimed at teenagers, describing the services our clinic has to offer.

As you read that last objective, if you wondered where you were going to distibute the 500 posters, then you are beginning to understand how to read and edit behavioral objectives. One could assume that you will distribute the posters in places where teenagers will see them, and not simply hand them to the first 500 people who pass the street in front of your clinic. But assumptions can be dangerous, particularly when a performance evaluation will depend upon everyone understanding the nature and extent of those assumptions.

If you wondered why the objectives specify 10,000 teenagers and 360 public service announcements, then you are correctly questioning how to arrive at the standards implied by the objectives (see 2.3).

Before you go on to the next section you might want to try a simple exercise. Write a goal which your education program would like to accomplish, or select a goal from a list you are currently using. Then attempt to write a behavioral objective related to the accomplishment of that goal. Next, reread your work and underline those words or phrases in the objective which clearly specify each of the following things: (1) who is responsible for doing: (2) what, (3) by when, (4) how, and (5) how one might observe that the first four things were in fact accomplished. All five of those elements do not have to be explicitly stated. In the example related to public service announcements on radio, it is sufficient to know that radio stations must keep written records of public service announcements, and that those records can be made available.

2.2.2 "OUTCOME" OBJECTIVES. Every educational event should be based upon a written list of objectives which specify what the participants will know or be able to do when they complete the session, course, or workshop. Such a list of "outcome" objectives will assist you in designing and evaluating the program. The objectives will suggest the content, the educational methodologies (film, lecture, activity...), and the sequence of events for the program.

Suppose that one of your objectives for this calendar year states: The education staff will offer five education sessions to groups of new

patients each week during 1981 (excluding sessions which fall on holidays) on Mondays, Wednesdays, and Fridays at 3:30 p.m. and Tuesdays and Thursdays at 5:30 p.m. Then you might write the following list of outcome objectives.

After completing the new-patient education session, the patient will be able to:

1. Identify, when asked to do so by an educator, the period when a fertile woman is most likely to become pregnant from intercourse during a normal menstrual cycle.

2. Identify the location and state the function of the following organs when shown a diagram of the male and female reproductive systems by an educator: testes, epididymi, vas defera, seminal vesicles, prostate, penis, vagina, clitoris, labia, cervix, uterus, Fallopian tubes, ovaries.

3. Make an appointment for his/her first clinic visit.

4. State the seven components of a Pap, pelvic, and breast exam when asked to do so by a clinic staff member.

5. State the correct instructions for preparing for a PP&B exam: (a) no douching for 48 hours prior to the exam; (b) no medications or contraceptive chemicals; (c) no intercourse for 24 hours prior to the exam; and (d) bathe no more than 24 hours before the exam.

6. List the following methods of contraception in descending order of relative use-effectiveness and state their modes of action, three advantages, and three disadvantages of each: the pill, the IUD, the diaphragm, the condom, vaginal spermicides, fertility awareness, and withdrawal.

A complete list might fill several pages, which raises the question of how much material one session can include. Some educators attempt to inform the patients and give them all of the reproductive health care information they will ever need in one hour! Obviously, some priority decisions will be absolutely necessary.

With this list of outcome objectives you can evaluate your education system to determine whether or not the patients are learning what you intend them to learn. And, you can compare different methods, and different systems, to see which does the best job. A list of outcome objectives can also help create lesson plans and help you train new-patient class instructors.

Before I proceed with objective writing in the more difficult areas, I want to include some thoughts on how to avoid being abused by objectives and how to avoid allowing the objectives to become the proverbial "tail which wags the dog."

2.2.3 PITFALLS.

There are several common mistakes in writing objectives and in using them to provide structure. Many beginners write objectives which are not objectives; if you cannot hear, see, touch, taste, or smell everything necessary to determine if the objective has been accomplished, your statement is not an objective.

Sometimes professionals write objectives covering the most insignificant things which they do as a part of their jobs (like showing up for work, opening the mail, taking coffee breaks, etc.). Writing objectives about such activities is a serious abuse of this system.

On the other hand, some beginners fail to include in their objectives activities which consume significant amounts of their time, taking for granted that those duties or tasks are understood. Writing grants, attending meetings (in-house, local, and out-of-town), traveling, interviewing "walk-ins," and answering correspondence can consume large amounts of time.

More frequently, beginners fail to reduce major tasks into several objectives representing stages of accomplishment. Even more often, some beginners attempt to accomplish a task or to complete an objective in one year which could easily consume two, three, or even more years. In the first case the problem is that a single objective needs to be reduced to several smaller objectives. In the second case the large project should be reduced into several sequential components, each represented by a set of objectives for one year. For example, suppose an educator has written the following objective:

During 1980 we will write and and print our own literature on four topic areas and have the material ready for distribution to patients by January 1, 1981:

1. our clinic procedures and services;
2. vaginitis;
3. Pap smears;
4. what to do if you think you are pregnant.

This major objective clearly specifies what will be done, by whom, when, and how. While the objective might appropriately appear in that form in a list submitted for formal approval, the following breakdown might help to clarify the steps. The educator responsible for the final product might write these objectives:

1. by May 15, 1980, the education staff will present clean, triple-spaced, typewritten copy consisting of 1,000 to 2,500 words on each of the four topics listed above to the Chairperson of the Education Committee and the Executive Director;

2. by June 15, 1980, the Education Committee will be prepared to meet with the Education Director and the Executive Director to discuss the content of the pamphlets and the changes which the committee will require before publication of the pamphlets;

3. by August 1, 1980, the Education Director will give the final copy of the text of each pamphlet to a printer along with camera-ready copy of any figures, charts, graphs, illustrations, or photos, and a dummy of the layout for each pamphlet;

. . . and so forth. One could, of course, divide each of these objectives into more sub-objectives, ad infinitum. Obviously, some judgment must be exercised.

You have also noticed, no doubt, that this process enables one to conduct what education experts frequently refer to as **task analysis.** That is, to specify in writing exactly what must be done and in what sequence, in order to accomplish a behavioral objective.[1] Experience has shown that this process is a useful, if not essential, part of accomplishing any complicated task, such as conducting a workshop or completing an edition of a newsletter (see 3.3.4).

2.2.4 SOME HARDER ONES

Sometimes professionals find it difficult to write behavioral objectives for activities or performances subject to the opinion of the observer

---

[1]Task analysis also refers to the process of listing every task which a departmental staff performs, however large or small the task may be, and then making logical allocations of the tasks to each staff member. This process proves valuable and time-saving when attempting reorganization (or initial organization).

in terms of their quality. For instance, family planning educators can "read" an audience and tell when they are holding attention and interest. When educators themselves sit in the audience, they can separate the good speakers from the poor ones. Yet, if some were asked to state how they formed their opinions they might have difficulty doing so.

Nonetheless, they form their opinions from what they hear, see, and otherwise perceive. Therefore, logic leads to the conclusion that those elements which comprise good public speaking can be observed, specified, and committed to paper. Similarly, other kinds of activities can be analyzed and their component parts specified.

Doing these analyses requires more time and effort than writing other kinds of objectives. First, you need to train your mind to analyze how the behaviors you observe combine to make the impressions you form. The process involves three steps:

1. Observing the person to be evaluated perform the task under analysis or observing one or more people whose ability to perform the task is considered good or strong.

2. Making a written record of those behaviors, those things perceived by the senses, which you notice about the performer.

3. Using those recorded observations to write lists of behaviors which enhance the performance and those which detract from it.

Nothing, however insignificant, should go unnoticed. The observer must focus her/his awareness solely upon what the senses record, and avoid drawing conclusions about the sensory input. Whenever possible, several persons should observe the performance and make records of observations, along with the time of each observation, to allow comparisons later.

Athletes use videotapes of their strides, swings, and strokes to identify mistakes and to maximize the potential of their performances. They analyze the components of the motion, examining each for flaws, and then practice to improve their weaknesses. Then, they integrate their lessons into a new and hopefully improved performance. What I am suggesting is that we educators use the same kind of approach to improve our skills and to train others.

Example one. Consider the "good public speaker." While this list is **not intended to be a prescription**, observing these things might suggest that a public speaker did an effective job.

BEST COPY AVAILABLE

The speaker is able to, or does:

1. Interview the person who invites the speaker at the time the initial request is made to ascertain exactly where the presentation is to be given, the exact date and time, the composition of the audience, the time period for which the speaker is invited to talk, the number of persons expected to be in attendance, and the topic or topics to be addressed.

2. Ascertain that all necessary equipment and instructional aids will be available when needed.

3. Arrive between five and ten minutes before the scheduled time to begin the talk.

4. Arrive bearing a personal appearance which is appropriate to the occasion and which presents a good impression of the agency which the speaker represents.

5. Begin his or her remarks with one or more of the following attention-getting techniques:

   A. five to ten seconds of silence;
   B. a statement which elicits laughter from one or more members of the audience;
   C. a rhetorical question.

6. State, within the first two minutes of his or her remarks, the purpose of the talk he or she is giving, any "ground rules" regarding when questions from the audience are appropriate, when intermissions or breaks, if any, will be taken, and the name and address of the agency which the speaker represents.

7. Maintain eye contact with members of the audience, regularly changing eye contact in a sweeping motion throughout the audience.

8. Be sufficiently familiar with the subject matter of the presentation that s/he refers to notes no more than three times every two minutes and no more than five seconds at a time, if at all.

9. Speak in a voice which varies in tone, volume, and inflection.

10. Speak sufficiently loudly to be clearly heard in every part of the room or by every member of the audience who is not hearing disabled.

11. Pronounce all words distinctly enough so that every member of the audience can discern each word.

12. Use the correct standard diction, syntax, and grammar of the native language of the majority of the audience.

13. Ingnore or control distractions or interruptions from within the audience and from outside of the room where the presentation is being given.

14. Stop the presentation to answer questions and then begin his or her material again without losing continuity.

15. Complete his or her remarks within the allotted time, including sufficient time for questions either during the presentation or at the end of the presentation (allowing time for one question for each five or six participants).

16. Refrain from using nonsense syllables (e.g. "umm," "uh," "errr") and extraneous words (e.g. "OK," "Ya see," "now then").

17. Thank the person or persons who invited the speaker for the opportunity to speak, either before the audience or in conversation following the presentation.

When you read the above list, if you noticed that items 4 and 9 could stand better behavioral analysis, then you understand the nature of these tasks. In item 4, for example, what is "appearance appropriate to the occasion" other than an individual's opinion? One could specify such things as "a speaker has good personal appearance when the speaker's hair is combed, when the speaker's clothes are clean, neatly pressed, and similar in style and mode of dress to the majority of the audience." However, even that slightly more behavioral description of the situation could stand further amplification.

In item 9, varying one's voice in terms of "tone, volume, and inflection" could mean something as simple as the speaker's voice is not a monotone, or something as complicated as the ability to deliver a Martin Luther King-type sermon. However, item 9 is sufficiently explicit to allow analysis of individual situations, case by case.

Example two. Another useful application of the techniques presented in this section involves analyzing an employee's performance in terms of behavioral objectives.

Suppose, for example, that one educator's basic problem in performing his or her work seems to be a gross lack of organization. However, it seems unfair, and may even be indefensible later, to simply say to the employee, "You are disorganized." What, then, constitutes an organized person? The following list of characteristics are things which organized people do or are capable of doing:

1. Locate specific items (books, pamphlets, letters, memos...) in, on, or around his or her desk in three minutes or less.

2. Make a daily list of tasks which s/he plans to perform during that work day in order of relative importance and according to deadlines established either by the employee or by someone else.

3. Prioritize tasks on a weekly, monthly, and quarterly basis.

4. Practice time management: plan all segments of his or her week as to when tasks will be performed based upon reasonably accurate estimates of the time each task will comsume.

5. List, in sequential order, those tasks which must be performed to accomplish or implement a lengthy or complicated project (e.g. budget and grant writing, evaluating, conducting workshops, lending educational materials ...).

6. Keep a readily accessible, alphabetically arranged record of names, titles, addresses, and telephone numbers of relevant professional contacts.

7. Keep an accurate, legible, and current calendar of professional engagements and commitments.

8. Arrive on time for meetings and other professional commitments and have in his or her possession all the materials necessary to complete the tasks at hand.

9. Accurately and understandably communicate with co-workers when jointly performing a task; accurately send and receive communications regarding who will perform which tasks, how, and by when.

Once again, this list is not a prescription and you could amplify some of the items in the above list by adding more items. Notwithstanding the possibility that the list might be lengthened or that the individual items might be embellished, the above list of nine items much more clearly communicates what it means to be organized than simply saying, "You need to get yourself organized."

2.3 RELATING OBJECTIVES TO REALITY

Now that you have learned how to write behavioral objectives, turn to the questions related to ensuring that the objectives will be meaningful and relevant indicators.

At first glance the ideas in the next few pages may seem obvious. **However, careful consideration of these ideas and attention to them during the objective-writing process is CRUCIAL to the application of this entire manual.** These ideas concern (a) the decisions made about the standards which the objectives set and (b) who will agree that those standards are realistic, relevant, and acceptable.

Social scientists, whether academicians or service providers, have argued for most of this century about behaviorism and other theories to explain human personality, thought, and action. The total extent of their debate has filled several lifetimes of study and numerous books and journals. With respect to evaluation the debate might sound like this fictitious debate between two evaluators:

Evaluator 1: "Behavioral objectives are indispensable to me. If I do not have a set of objectives against which I can compare evaluative data, then anyone can interpret the data to mean anything s/he wishes! But with objectives, I know what data I will need and I can draw clear conclusions about what the data mean."

Evaluator 2: "You just want a crutch! Evaluation by its very nature is subjective, the assigning of value judgments to data. If you write 'behavioral objectives' you simply set arbitrary standards. Then you collect your data and say, 'Well, we met these objectives and we didn't meet those.' So what?"

1: "But what alternative do you offer? You would have evaluators apply arbitrary and, I might add, capricious, standards to the data once you have collected it. No wonder evaluation has such a bad name! You would have every educator place his or her performance ratings in the hands of some stranger, based upon blind faith!"

2: "Perhaps. But all you want to do is move the arbitrary standard-setting process from the end of the evaluation process to the beginning--and 'sanctify' it by writing 'objectives.' I still say, so what?"

1: "But values are always arbitrary. When we specify the values we intend to use at the outset everyone at least knows which values, whose values, will be applied to the data at the end."

2: "And what do we do when the values we set at the beginning turn out to be meaningless at the end? You know as well as I do that things always arise someplace along the line which no one foresaw or could have foreseen. Every educational program has some unintentional effects."

1: "Naturally. I am not advocating that we ignore those additional effects. But I want to set objectives so that the program as a whole can be evaluated against some set criteria."

2. "Which brings you right back to my philosophy! You have to write explanations for why you met your 'objectives' or why you didn't. Which is nothing more than compiling subjective opinions. Which is what I do, but with some judgments, not meaningless 'objectives.'"

This debate will rage on for some time to come, primarily because both sides have valid arguments, and both have important counterpoints to the other's ideas (see, e.g. Scriven, et al., 1975).

Objectives are valuable tools if used as such--as a means to an end rather than an end in themselves. Objectives are a means of setting standards against which to compare evaluative data. If the objectives are

written with arbitrary standards in mind then Evaluator 2 is correct; all you do is move the process of being arbitrary from the end of the process to the beginning and lose the flexibility you may need later. On the other hand, while setting standards will always be somewhat arbitrary, that process should not be capricious or irrational. Therein lies the crux of the matter--objectives written without rational consideration of their meaning and purpose constitute an abuse of the system and of those who work toward them and evaluate against them.

There are many instances when family planning educators can write meaningful objectives against which to evaluate their work. The standards which objectives set can be rational, based upon logical analysis of the goals upon which the objectives are based and upon reasonable criteria. For instance, consider the problem of answering requests for speaking engagements about family planning. Should an education staff fulfill all of the requests they receive? Should they set an objective of soliciting a certain number? Should they solicit ten? Forty-two? Two hundred? Obviously, some other data must come under consideration.

Standards for objectives may be set in terms of **expectations compared to background data.** If you fulfilled 425 requests last year then you might decide to increase that by five percent or ten percent. However, arbitrary choices of a certain percentage will not suffice either. You need to **consider the allocation of resources** as a standard for setting objectives. How much time did it take to fulfill 425 requests last year? Based upon the relative priority of the populations requesting presentations, and upon the other projects you have in mind, how much time can you devote to this aspect of your work? (More information on timing and budgeting appears in 3.3.4).

Sometimes an objective may be set against a **norm,** some characteristic of a group or some data collected in prior research. You might set an objective for reducing teenage pregnancy against vital statistics or school-system data on student pregnancies. You might establish training objectives based upon data which reveal the nature of the skills and knowledge needed to perform a certain set of tasks. Norms may also be established using written tests **(see section six).**

Rational standards may also be set in objectives at the **mastery level,** i.e. the knowledge and skill necessary to perform some task, make some

decision, complete some process. A new-patient education class might be evaluated against objectives which specify what patients need to know or be able to do in order to be effective contraceptors, or knowledgeable clinic patients (see 2.2.2).

Sometimes **external standards** are set by other organizations or government agencies. Informed consent requirements and content for patient education are specified in most instances by medical practices, government agencies, or national family planning organizations. Sometimes these external standards are arbitrary and basically meaningless, needing further refinement. In some cases they have been carefully developed and logically prescribed.

Professional judgment is a valid way to set standards. For example, your experience and expertise, or that of your peers, might set valid standards for training volunteer speakers. By valid standards I mean that the standards are defensible in terms of what you would expect the volunteers to do after their training, which would easily lead you to "outcome" objectives for the training.

When none of the above methods for setting standards applies, or no data are available to set standards, then perhaps you need to consider gathering background data in order to set norms in an educational activity. Or you might convene a panel of experts to establish standards through professional judgment.

In some instances we cannot set meaningful standards for educational activities through behavioral objectives without making certain assumptions. Then, the validity of the standards set by the objectives depends upon the assumptions. **Attitudinal measures** and work aimed at changing or assessing participants' attitudes offers an excellent example. The currently available measures of attitude in any affective area, whether sexuality or some other aspect of life, rely either upon interpretation of written test scores or upon physiological measures, such as attaching electrodes to a subject's skin. The written tests are the subject of an ongoing debate as to their validity, and the physiological measures have such an effect upon the subject that their intrusion into the attitudinal domain is clouded by the distractions of the wiring. If you assume that one of these measures will accurately measure attitudes then you can set objectives using standards against the test results.

In other instances, no reliable, valid, and rational standard can be set. In those cases, we are only being professional and ethical when we state that plainly and proceed accordingly, revealing the assumptions which we are making and carefully listing the limitations upon the data we collect and their interpretation.

Ultimately, the validity of the standards which we set should be determined by those who will use the evaluative data. At this point, the educators, executive director, and the education committee should agree upon the content of the objectives and the standards which they imply, and upon the data which will be collected to evaluate the performance of the education program against the objectives. As indicated in the discussion of external standards, agencies such as health systems agencies (HSAs), regional offices of the US Department of Health and Health Services (US DH & HS), and regional, statewide, or national organizations may set criteria for evaluating family planning education programs or activities. These criteria are of obvious importance in the objective-writing and standard-setting processes. The roles and functions of these outside evaluators need careful delineation, and their requirements should be clearly specified in advance whenever possible.

Every objective should be subjected to a "So What" Test: ask yourself, "If I accomplish this objective, so what?" The answer should clearly relate to the mission of the education department, not to some need to have everything covered by objectives. Subsection 3.3.2 suggests that the objectives must relate to their goals. Here, the "So What" Test refers to the validity of the standard set by each objective as a measure of accomplishment; does the standard have supporting, defensible logic, or is it an arbitrary and capricious "mark on the wall"?

At the expense of overemphasis, objectives are **tools.** When writing them, rewriting them, and reporting on their accomplishment becomes an end, or **the** end, the people involved are abusing objectives and allowing themselves to be abused by the process. While written objectives cannot replace evaluative judgment, they can provide the magic of structure.

Section three explicitly describes the planning and evaluation process for a year's worth of programs or for a single educational event. The subsequent sections contain specific technical advice on the methods you will need to use for collecting evaluative data. **Section ten: A THOROUGH**

EXAMPLE includes an evaluation strategy for a hypothetical education program and sample evaluation questionnaires.

You probably have a great deal of consulting help available from universities, health systems agencies, national and statewide family planning organizations, and from colleagues and peers. If you work for an affiliate of the Planned Parenthood Federation of America, help is as near as a telephone call to the Department of Education at (212) 541-7800.

SELF-TEST ON SECTION TWO

This is a self-test on the material in **section two.** All of the answers to the questions in the self-test are contained within this section. The correct answers may be found on the page following the end of the self-test. You will probably learn more, and get a much better idea of how much you have learned, if you write down your answers on a separate sheet of paper and then check them after you have completed the self-test.

1. What is a list of priorities for an education program?

2. What is the difference between a goal and a behavioral objective?

3. (a) Write a list of four priorities.

   (b) For the first priority in (a) above, write a goal.

   (c) Write three behavioral objectives for the goal you wrote in (b) above.

4. In 2.2.4 you were given examples of criteria which defined an "organized" employee. Add three criteria to that list.

5. One of the following statements is a behavioral objective and the other four are not. Identify the one which is a behavioral objective and state why each of the other four statements is not a behavioral objective.

   (a) A one-page instructional handout will be written by April 1, 1981, at the seventh-grade reading level, describing the Pap smear and covering the reasons why women need a Pap smear, what a pelvic exam entails, and what a patient should do to prepare for a pelvic exam which includes a Pap smear.

   (b) The Education Director will report on the evaluation of the school nurses' workshop within three weeks of the date when the workshop ends (May 1, 1981).

   (c) The Education Secretary will purchase five books on childbirth which are not already in the affiliate's library collection, prepare them for placement in the circulating collection, and place them on the shelves ready for circulation by July 15, 1982.

   (d) The Public Relations Associate will prepare a press release on our new fertility-awareness program written in the Associated Press style and distribute it to both local newspapers and all seven local radio and television stations.

   (e) During 1981 the education staff will distribute 5,000 copies of our new pamphlet on teenage pregnancy.

6. Ideally, who are the individuals or groups who should participate in the process of setting priorities, writing goals, and writing objectives, and in what order do those individuals participate in that process?

7. List two arguments against writing and evaluating against behavioral objectives and two arguments in favor of this approach.

8. State a principle to guide you in the setting of standards implied by behavioral objectives.

9. Add three "outcome" objectives to the lists presented in the example on page 18.

ANSWERS TO SECTION TWO SELF-TEST

1. A list of priorities for an education program is a list of either groups of persons or concerns/problems, written in order of relative importance, to be used for the purpose of allocating educational program efforts, funds, and other resources.

2. A goal is a global statement of an educational program's desired outcome, whereas an objective, or behavioral objective, is a specific declarative statement which defines who will do what, how they will do it, where and when it will be done, and how one will know when it has been accomplished. The evaluation of whether or not you have reached a goal may be a matter for subjective judgment; an objective, however, can be judged as having been accomplished or not through data which are perceivable through the senses.

3. There are a myriad of answers to this question. Examples are given earlier in this section. You have correctly performed the tasks asked for in question 3 if your answers meet the definitions given in answers to questions 1 and 2.

4. There are a variety of possible correct additions to the list. Your answers are correct if they relate to organizing a professional's performance of tasks **and** if they are statements which allow an observer to ascertain whether or not a person is or is not doing what you have written through the **senses.**

5. (a) This is not a behavioral objective because it does not specify who will write the pamphlet on Pap smears.

   (b) This is not a behavioral objective because, although the objective states that the Education Director will make a report, it does not state in what form the report will be made or to whom the report will be transmitted.

   (c) This is a written behavioral objective.

   (d) This is not a behavioral objective because it does not state a time when the Public Relations Associate will have written or distributed the press release.

   (e) This is not a behavioral objective because it does not specify what methods the education staff will use to distribute the new pamphlet on teenage pregnancy or to what groups or individuals the pamphlet will be given.

6. Ideally, the education staff starts the process of writing priorities, goals , and objectives by writing a list of priorities, and then lists of goals for each of the priorities. Next, the education director and chief administrative officer of the family planning agency (or the administrative staff member responsible for planning and evaluation) discuss and agree upon the content of the lists of priorities and objectives. Then, in sequential order, the education committee and the entire board of directors review and comment upon the lists of priori-

ties and goals. The lists of priorities, goals, and objectives are distributed next to other professionals and to representatives of the target populations which the education program will serve for review and comment. Finally, the education staff, in cooperation with other agency staff who may be involved in accomplishing the goals, write behavioral objectives for each of the goals.

7. A. (1) Behavioral objectives require too much time to write them and to discuss; (2) behavioral objectives are arbitrary and meaningless standards against which to evaluate; (3) writing objectives merely moves the arbitrary process of setting standards from the end to the beginning of the evaluation process; (4) evaluating against objectives leads to a false sense of satisfaction when meaningless standards are met; (5) objectives lead evaluators to overlook important but unintended or unforeseen outcomes.

   B. (1) When the nature of success and failure are nebulous, objectives are the only method of creating explicit standards and statements of expectations; (2) while some arbitrariness will always appear in any standard, objectives can lead to logical thought about standards, and stimulate rational approaches to creating standards in advance; (3) setting written objectives explicitly clarifies expectations and assists planning.

8. Standards set by written objectives should be based upon logical attention to prior performance, resource allocation, norms, mastery levels of knowledge or skill, or to logical reasoning based upon a clear statement that sufficient data do not exist to permit standards to be set by one of the previously mentioned methods.

9. After the new-patient education session, the patient will be able to:

   A. Describe the symptoms of vaginitis when asked to do so by a clinician.

   B. Identify the method(s) of contraception which s/he will use in the period between the new-patient class and the time when his/her permanent method of contraception will take effect when asked to do so by an educator.

   C. Name three services offered by this agency other than education and birth control methods when asked to do so by a staff member.

   Of course, there are many other kinds of outcome objectives which you could include in your list, depending upon the decisions which you make about your new-patient education system.

section three:
THE PLANNING AND EVALUATION PROCESS

This section is the skeleton of this manual, giving all of the other sections a structure and frame of reference. Beginners may find the prescriptions in this section to be a good background against which they can create their own approaches, adapting the theories of varied disciplines to the ever-widening experiences of family planning educators. Experienced educators, planners, and evaluators will find few novel constructs in this section, but they will perhaps find a different and thought-inspiring perspective on the planning and evaluation process.

After reading this section and learning the material in it you will:

1. Be familiar with an eight-step process which details the sequence described as the Cyclical Evaluation Model in **section one (Figure 1).**

2. Have a clear conceptualization of a needs-assessment process.

3. Be familiar with some of the arguments proposing and opposing the use of behavioral objectives as standards against which to evaluate.

4. Have principles to guide the process of setting standards for evaluating family planning education during the objective-writing phases.

5. Be able to identify an educational event's content, process, outcome, and impact, and have guidance as to when to evaluate all or each.

6. Have a familiarity with the basic vocabulary of data-collection techniques and be aware of your choices for collecting evaluative data.

7. Be aware of sources of information from the various libraries which collect materials related to family planning and sexuality education.

8. Be able to perform task analyses, cost analyses, and calculate cost-effectiveness of educational products and services.

9. Have explicit guidance for making decisions about whether or not to write an evaluation report in various instances.

10. Know of potential uses and sources of voluntary consulting help.

A self-test to aid in reviewing or to be used as a posttest appears at the end of this section.

section three: THE PLANNING AND EVALUATION PROCESS

> "If the only tool one has is a hammer, then one tends to treat everything as if it were a nail."
>
> --A. H. Maslow

The eight-step process outlined on the next page amplifies the Cyclical Evaluation Model presented in **Figure 1 (Section one).**

The eight steps bridge the gap between the process of writing goals and objectives and the data-collection methods needed to verify that the objectives have been accomplished. The remaining seven sections of **The Magic of Structure** contain the material which you will need in the data-collection process.

The eight-step process will work equally well for a year's worth of educational activities or for a single educational event lasting an hour. The process will help you document the effectiveness of your work, plan your time, allocate your resources, predict program costs, and apply what you learn about your programs to future planning.

Before describing the steps one by one, I want to attend to the question of what to evaluate in relation to any educational event--speech, workshop, course, ad campaign, seminar, and so forth. Any educational event has four aspects which may warrant specific evaluative attention: content, process, outcome, and impact. Educators may want or need to evaluate aspects of a given event or series of events that are different from aspects that administrators or board members care to evaluate.

3.1 WHAT TO EVALUATE

Content. The content of an educational event refers to the specific information or skills the educators want the participants to learn or improve during the program, or the attitudes to be changed. In many cases, the content of an event may be so obvious, or even so well prescribed, that evaluating content is unnecessary. For example, US DH&HS has specified what new family planning patients must be told by programs which receive federal family planning monies. They even specify that the content of the new-patient education session must be documented in each new patient's medical chart for patients aged nineteen and under.

OUTLINE OF THE PLANNING AND EVALUATION PROCESS

This outline presents an overview of the remainder of this section. The numbers in parentheses refer to the subdivisions of this section where each step is explained in detail.

I. Assessment Phase

Step 1. Assess the needs of the community or population at which you are aiming; write goals and priorities. (3.2)

II. Planning Phase

Step 2. Write behavioral objectives based upon the goals. (3.3.1)

Step 3. Ascertain that the accomplishment of each objective will be a step toward reaching the goal for which it was written. (3.3.2)

Step 4. Identify what kind of data you can collect to ascertain that you have reached your objectives. (3.3.3)

Step 5. Perform task and cost analyses of the steps which must be taken to implement the program and its evaluation. (3.3.4)

III. Implementation

Step 6. Implement the various tasks that accomplish the program and its evaluation. (3.4)

IV. Evaluation

Step. 7. Gather and organize the necessary data and calculate any statistics needed. (3.5.1)

Step. 8 Compare the data to the criteria established in the objectives, determine if the objectives were met or not, and report in writing to the appropriate person(s). (3.5.2)

In many other instances the content of an educational program is not subjected to evaluation, although it might be and perhaps should be. For instance, the content of a high school speaking presentation may be agreed upon by the teacher and the family planning educator. However, an educator or an executive director might decide that evaluation of the content, in terms of exactly what was said by whom, might be necessary for accounting to either the board of directors, the school district, or segments of the

community through the media. This specific example is treated in more detail in **section ten.**

Process. This term refers to the methods, activities, and chronological sequence used to present content. You consider educational process when you decide things like: In what order are we going to present these six topics? Are we going to present the information on birth control with a lecture or a film? Are we going to teach values clarification by lecturing, by doing values-clarification exercises, or neither? Are we going to do this sixteen-hour workshop in two eight-hour days, or four four-hour sessions on weekday evenings?

Process evaluations usually focus upon whether or not a given process is the best way to impart certain content, or whether a given process is the best way to get the desired outcome or impact from a program. For example, suppose you have agreed with your clinic staff about what kinds of things you want new patients to know when they arrive at the clinic for their initial pelvic examination. You have written behavioral outcome objectives specifying what you want new patients to be able to say or do when they arrive at the clinic. You are evaluating the process of your new-patient education system when you determine whether or not your new patients actually arrive at the clinic with that information. You already know what information they are supposed to have; if they don't have it then you need to change your educational methods. You are also evaluating process when you compare one or more methods, such as lecture versus film, to impart content on a specific topic. The research project example in **section eight** evaluates process.

People outside an agency, and even agencies' boards of directors, are frequently not concerned with the process of an educational event, but rather with its content, its outcome, or its impact. However, family planning educators need to be aware of the processes they are using at all times, in order to ensure that they are having the desired outcome and impact.

Outcome. The outcome of an educational event refers to what changes that program caused among its participants. You are evaluating outcome when you examine how the participants in the program differ immediately after the program from when the program started.

Outcome might concern cognitive, affective, or psychomotor dimensions, a combination of two, or all three. **Cognitive** changes refer to those changes regarding what participants know, or skills which they need in their daily lives. **Affective** changes deal with changes in attitudes or feelings, such as changing peoples' attitudes from less accepting of teenagers' using contraception to more-accepting attitudes. **Psychomotor** refers to specific skills which involve coordinating movements of the body, such as filling a foam applicator or inserting a diaphragm.

Impact. Impact refers to whether or not the outcome of a program lasted for some period of time. The dividing line between outcome evaluations and impact evaluations is always quite arbitrary. Clearly, you are evaluating outcome when you evaluate the participants and the changes which occurred in them immediately following an educational program. Just as obviously, if you wait a year to evaluate those things then you are evaluating impact. Impact is **also used to refer** to whether or not an educational program changed the lives or the behavior of the participants in some significant way. Sometimes an investigation of an agency's effects upon a community in terms of the problems the agency was created to alleviate is called an impact evaluation.

As you apply the planning and evaluation process to any given educational event, remember that **you may need to use different methods and tools to evaluate these different aspects of the event.** For instance, you may need to use a test to evaluate what the participants in a workshop learned, a questionnaire to gather their perceptions of the process you used, and still another approach to evaluate the impact which your workshop had upon the participants' ability to perform some kind of task. With respect to your patient-education system you may (1) evaluate your process to determine whether it taught the patients the required content,(2) evaluate the outcome by investigating their experiences when they attend their first appointment in your clinic, and (3) evaluate the impact by conducting a follow-up survey to determine whether or not they avoided unintended pregnancy and STDs, or to get feedback on whatever else you hoped to accomplish.

3.2. THE ASSESSMENT PHASE (Step 1)

A needs assessment may serve two basic functions. It may gather data to determine which groups within a community would benefit most from what

kinds of educational programming. Second, a needs assessment may provide data to guide decisions about the content, process, outcome, and desired impact of a given educational event. In the first instance, the needs assessment assists you in long- or short-range planning, in the second case with program planning.

All too frequently a needs assessment for a given educational event consists of an educator sitting with feet upon the desk thinking, "Well, let's see. From my past experiences with this kind of audience they need to hear about this, this, and that. Oh, and I think I'll take a film...." Long-range planning, and the needs assessment part of the planning process, is the most crucial and yet most often omitted step.

Step 1: Assess the needs of the community or population at which you are aiming; write priorities and goals.

Long-range planning. When you attempt to set forth your plans for educational efforts over the next three to five years or more, you are doing long-range planning. That kind of process should include a needs assessment of the entire community you plan to serve--a major undertaking which may require that you follow steps 2 through 8 below. Certainly, you should write some objectives which you intend to accomplish in creating your long-range plan.

A long-range plan should consist of no more than priorities and goals for the future. You can use those to create smaller lists of goals and objectives for your subsequent short-range plans, covering periods of time up to a year.

Begin with two important kinds of data. The characteristics of the population you intend to serve will offer excellent guidance: ethnic and cultural composition, age profiles, births to single mothers, public health data on STDs, family-unit characteristics, economic status and trends, and so forth. While you might rarely be able to conduct an extensive survey of the community to gather such data, you will probably have to rely upon the sources of data listed in **section four.**

Second, gather information about the other kinds of programs which will serve your community in family planning, family life, and sexuality education. You might be wise **not** to assume that your familiarity with the

other agencies and educators in the community completely encompasses all planned events, programs, and efforts. Instead, interview your peers, board members, school district personnel, health department staff, and others who may aim programs at the same groups for the same reasons. Ask for copies of other agencies' long-range plans.

Then, combine these two sets of facts to answer the following kinds of questions: Which groups most need family planning education? Which groups have been most seriously neglected by education in the past? How can my agency most efficiently use its limited resources to have the greatest impact upon the most needy groups? Your agency's mission statement should provide you with the ultimate, ideal impact. What goals should you set for your educational programs over the next three to five years to move closer to the accomplishment of the mission statement?

A long-range plan is actually a statement of agency policy and, as such, the responsibility of the board of directors. However, staff members should work closely with the board and its committees to provide professional expertise and start the communication process which will lead to the short-range plans and specific objectives.

Short-range planning. With the long-range plan completed, you can begin to take a harder look at the data gathered to write it so that you can prioritize the sequential steps toward accomplishing your long-range plans. The short-range planning process, yielding priorities, goals, and objectives for a year or less, may require that you return to the data base you used for the long-range plans and seek additional, more-specific data.

Section 2.1 describes the short-range planning process in detail.

Program planning. If you are conducting a needs assessment to ascertain what you should include in a specific program, such as an often-repeated lecture, a workshop, or a course, you will get the most-revealing data by contacting your audience directly. You might interview a group of them, survey them (see section seven), examine data from previous evaluations of the same kind of program, or use any of the data-collection strategies mentioned under step 4 (3.3.3).

When direct contact with the audience is impossible or impractical, you might interview or survey their leaders or the professionals who serve them. An on-site visit to meet with staff, clients, or both often proves revealing if a clear agenda is established and completed.

When another agency or organization offers a contract to perform an educational service, such as a training workshop, you might be wise to assess the needs of the audience in addition to any content or process prescribed by the contractor. For instance, if a local school district requests a workshop on basic family planning information for its nurses, you would be wise to survey, or interview, or use both approaches, to assess the needs of your potential participants. A needs assessment survey questionnaire jointly constructed by the administrator for nurses and by the educators conducting the workshop should satisfy all concerned.

3.3. THE PLANNING PHASE (Steps 2 - 5)

Four of the eight steps in the planning and evaluation process occur during the planning phase. With a sound needs assessment you are ready to devise specific strategies to meet the needs of your community and your individual audiences.

3.3.1 Step 2: Write behavioral objectives based upon the goals.

Since section two describes the process of writing behavioral objectives, I will not repeat that information here.

However, if you did not read section two because you are familiar with writing behavioral objectives, I would urge you to read section 2.3 which **contains information essential to the correct application of this process to educational programs.**

If you are writing short-range planning objectives, the process is outlined in section 2.1. If you are executing a program plan for a specific educational event, see 2.2.2, "outcome" objectives.

3.3.2 Step 3: Ascertain that the accomplishment of each objective will be a step toward reaching the goal for which it was written.

All helping professionals, whether they work in service-delivery systems or in academic communities, must make certain assumptions about the work which they do. This step should reduce the number of assumptions which we make about family planning education and its outcomes and impacts. Sometimes what seems intuitively obvious turns out to be contrary to reality. Can we assume that telling high school sophomores about the locations of our clinics and the kinds of services we offer will motivate more of them to come to our clinics for contraceptives? Or, that a thirty-

minute talk about contraceptives to a group of new patients will give them all of the information they need to make rational choices about which contraception each will use? Can we assume that a demonstration of empathic listening will improve trainees' empathic skills?

Family planning educators make these kinds of assumptions--and some even bigger ones--daily. Some of these assumptions may be valid for one type of content and process in an educational session, but invalid for another. To complicate the situation even further, one educator might have the desired impact with a specific content-outline and process but another educator might not achieve the same ends with the same methods.

I believe that we family planning educators need to spend more time and effort evaluating the efficacy of what we do, and more time ascertaining that our assumptions are true. If we go through the process of setting a goal, writing objectives aimed at accomplishing that goal, collecting data to verify that we accomplished our objectives, and evaluating what we have done by comparing the data to the objectives, **WE HAVE NOT GOTTEN A STEP CLOSER TO REACHING OUR GOAL UNLESS THE OBJECTIVE ITSELF REALLY MOVES US CLOSER TO THE GOAL.** After you have written your objectives, ask yourself the following question about each objective: "How do I know **for a fact** that if I accomplish this objective I will be getting closer to my goal?" If you cannot give a specific answer to that question, preferably in a behavioral form, then it is important that you evaluate outcome, impact, or both, to ascertain that you have gotten closer to your goal. The results of that evaluation may indicate that you need to return to the beginning of the process and write objectives aimed at the relationship between the impact of an educational program and the attainment of a goal you have set.

For example, suppose you have the following goal for your education program:

GOAL: We will increase the use of contraception by sexually active teenagers in our community.

One of the objectives you have written for that goal states:

OBJECTIVE: We will speak to a total of 10,000 teenagers in schools, churches, clubs, and other youth-serving agencies during 1981.

It seems to me that we family planning educators have been assuming, perhaps for decades, that speaking to more and more teenage audiences will increase their use of birth control if and when they are sexually active. Is that a valid assumption?

Teenagers, in order to avoid pregnancy, must obviously either not have sex or use contraception. The question remains as to whether talking to teenagers about our services, their values, their choices, contraceptive methods, or a combination of those things will motivate them to start using contraception when and if they have sex. Among those teenagers who do have sexual intercourse, several things must happen between the time they hear a family planning educator speak and the time they begin to use contraception. First, they have to make a conscious decision to obtain contraceptives. Then they have to go to a drug store, a physician's office, or a family planning clinic and actually obtain the contraceptives--and use them--consistently and conscientiously.

If we simply assume that talking to more and more teenagers will reduce the number of teenage pregnancies without verifying that that will work, we may be spending a great deal of time and effort reaching our objectives, but not move a step closer to our final goal!

There are various ways to ensure that your assumptions about the relationship between accomplishing an objective and reaching a goal are valid. Sometimes the research or evaluation published by others will offer sufficient evidence. For example, from various sources we know that handing a new patient, particularly a poorly educated or illiterate one, a package of birth control pills without explicit instruction for use will not help prevent pregnancy. We know that people remember more about a specific content-area when they see instructional aides, such as audio-visuals and tangible objects, than when they simply hear a lecture about something. They remember twice as much of what they read as what they hear.

In essence, step 3 refers to assumptions about cause and effect: does an A cause C, or does some other intervening event or process, B, really lead to C. In family planning and sexuality education, there are likely to be several B's between A and C.

Perhaps family planning educators should commit their evaluative evidence to paper and share it with others, to avoid duplication of effort and to accrue a convincing body of evidence of effectiveness.

3.3.3 Step 4: Identify what kinds of data you can collect to ascertain that you have reached your objectives.

This section's opening quote was chosen with this step in mind. Devising sound strategies to collect evaluative data is important for several reasons. First, if the data-collection process itself consumes too much time, we cease to be educators who evaluate their work and become evaluators who do a little education now and then. Second, without the proper data in a usable form collected in a logically defensible fashion, the objectives you have written become virtually meaningless because you cannot know whether you accomplished them or not. Finally, evaluation interfaces with research most closely at this step. Not every problem, question, or evaluation objective lends itself to sound data-collection strategies because the science of research in the helping professions lags far behind its counterparts in the physical and life sciences. There are, however, many good methods available. But you cannot use these tools unless you know how to find them. You do not need to become an expert in research design and statistics to use them, but you do need to have a basic familiarity with the language and process in order to communicate with experts who do have the necessary knowledge.

All of the following techniques are covered elsewhere in this manual: behavioral observations **(section eight)**, descriptive statistics **(section nine)**, experimental research **(eight)**, questionnaires (five), **other people's** data **(four)**, surveys **(seven)**, and tests **(six)**.

Do not overlook libraries! A review of the relevant literature can provide invaluable information throughout all four phases of the planning and evaluation process. The recent advent of computer searches has reduced the boredom and tedium of library research. For a relatively small fee, usually ten to fifteen dollars, academic and special libraries can conduct a search of the publications in several fields to find what you may want.

The following libraries or institutions have collections directly related to family planning and sexuality education and often produce bibliographies on topics related to our work:

SELECTED LIST OF CENTERS PROVIDING RESOURCES

(Kinsey) Institute for Sex Research
Indiana University
416 Morrison Hall
Bloomington, IN 47405
(812) 337-7686

International Clearinghouse
for Family Planning Information
P.O. Box 2225
Rockville, MD 20852
(301) 881-9400

Institute for Family Research
and Education
Syracuse University
760 Ostrom Avenue
Syracuse, NY 13210
(315) 475-8762

Population Information Program
Johns Hopkins University
624 N. Broadway
Baltimore, MD 21205
(301) 955-8200

Hopkins Population Center
Hopkins School of Hygiene
Public Health
615 N. Wolfe Street
Baltimore, MD 21205

Learning Resources Center
Life Planning/Health Services
2727 Oak Lawn, Suite 228
Dallas, TX 75219
(214) 522-0290

Harvard Center for
Population Studies
665 Huntington Avenue
Boston, MA 02115
(617) 732-1234

Sex Information and Education
Council of the US (SIECUS)
84 Fifth Avenue, Suite 407
New York, NY 10011
LIBRARY: 51 W. 4th St. Rm. 53
New York, NY 10003
(212) 673-3850

Center for Disease Control
US Public Health Service
Family Planning Evaluation Division
Building 1, Room 4403
Atlanta, GA 30333
(404) 329-3995

Center for Population and
Family Health
Columbia University
60 Haven Avenue
New York, NY 10032
(212) 694-6985

Urban Institute
2100 M Street, NW
Washington, DC 20037
(202) 223-1940

Center for Demography
University of Wisconsin
3216 Social Science Building
Madison, WI 53706
(608) 262-2182

Carolina Population Center Library
University Square
Chapel Hill, NC 27514
(919) 933-3006

The Clearinghouse for Sexuality
and Population Education
Program Materials
Department of Education
Planned Parenthood Federation
of America
810 Seventh Avenue
New York, NY 10019
(212) 541-7800

Katherine Dexter McCormick
Library
Planned Parenthood Federation
of America
810 Seventh Avenue
New York, NY 10019
(212) 541-7800

Association for Voluntary
Sterilization
708 Third Avenue
New York, NY 10017
(212) 573-8377

3.3.4 Step 5: Perform task and cost analyses of the steps which must be taken to implement the program and its evaluation.

The process of forseeing each sequential step involved in implementing **and evaluating** a particular educational program, or the various programs which you will do during a week, month, or year is a **task analysis.** Each of those tasks will involve certain expenditures of time, resources, and other costs, a prediction of which is a **cost analysis.** A complete cost analysis of the various programs included in a year would easily permit estimating a budget for the year.

Since both of these processes require prediction, some inaccuracy is inevitable. Nonetheless, a written list of task assignments for each person involved in implementing an educational program immeasurably increases the probability that important tasks will get done in sequence and on time. A cost analysis will provide raw data not only for compiling budgets but also for determining **cost-effectiveness,** the cost to produce a given unit of outcome or impact.

Some educators, particularly those with a great deal of experience, perform task analyses mentally, never committing the list of tasks to paper. However, even the most experienced educators often find a written list of tasks invaluable when they collaborate with others to implement a program. At minimum, there needs to be a clear and explicit understanding of who will be responsible for doing what when. A task analysis does not have to be a formal, neat, type-written document; simple, handwritten lists should suffice.

However, a task analysis is a prerequisite to a cost analysis. With the task analysis you can estimate the number of hours and other costs involved in completing a given set of tasks. The following task analysis for producing one flyer shows estimated person-hours:

	ESTIMATED HOURS FOR:				
	Educators				
	A	B	Artist	Typist	Totals
TASKS:					
Draft the text	8				8
Type draft				1	1
Review and edit	2	3			5
Layout art design			2		2
Type final copy				2	2
Prepare camera-ready art			4		4
Prepare dummy of art & text			2	1	3
Printing	2	2			4
Retrieve, review, & return proof	2	2			4
Receive shipment, pay invoice				1/2	1/2
TOTALS	14	7	8	4 1/2	33 1/2

With the above information you can calculate the total personnel costs by multiplying each person's hourly wage (including fringe benefits) by his or her estimated number of hours and summing the results. The remaining costs of this project would include the printer's charges for printing the flyer, the mileage expenses to travel to and from the printer, artist's supplies, and office supplies for the typist and educators. With that data in hand, you can calculate the total cost of the project to produce the flyers, and the unit cost to produce one flyer (the total cost divided by the number of flyers printed). If you produce 3,000 copies of the flyer at a total cost of $415, then each flyer costs $415.00 ÷ 3,000 or about 13.8 cents each, the cost-effectiveness of producing that flyer in that manner.

If you perform task and cost analyses for each objective you plan to accomplish during a three-month, six-month, or twelve-month period, you will have a good start toward a budget for that period. A similar process will yield a budget estimate for a grant proposal. You will also have a data base for allocating staff time over the same period.

Budgeting. **Figure 3,** a checklist of items to remember when analyzing projects, should ease your cost analyses of individual projects. **Figure 3** also includes some items which appear on annual budget sheets but which do not necessarily appear in the cost analyses for individual projects, or which do appear but need additional analyses for annual budgets. For example, local telephone charges, rent, and insurance costs do not typically appear on the budget for an individual educational event but will always appear in an annual budget.

Figure 3. Budget Planning Checklist

The following items might appear on the budget for an individual education program, in addition to the personnel costs.

_____consultants/guest speakers

FACILITIES
_____meeting room rentals
_____meals served to participants
_____coffee, tea, snacks

EDUCATIONAL MATERIALS
_____pamphlets
_____books
_____film rentals
_____folders for materials
_____other audio-visual rentals

TRAVEL
_____local mileage, staff & others
_____out-of-town transportation
_____lodging, meals, incidentals

OFFICE SUPPLIES
_____stationery
_____pens, pencils, tape, markers
_____newsprint pads, rolls
_____typewriter supplies
_____plain paper, index cards, etc.
_____fasteners, staples
_____large envelopes

MISCELLANEOUS
_____postage, handling, shipping
_____telephone, long distance
_____printing
_____photocopying
_____security services

In addition to the items above, you might include the following in your budget projections or analyses for an annual budget. Do not forget that some of the expenses above might be more for a year than the sum of your analyses for each objective because of "overhead" expenses or routine tasks which you do not cover in your objectives.

_____occupancy (rent, mortgage, utilities, insurance, garbage collection, fire protection, security guard services)
_____telephone, both local and long distance
_____equipment (typewriters, copy machines, adding machines, projectors, duplicators, furniture,...)
_____library books
_____library supplies (book pockets, paste, catalogue cards, spine labels,...)
_____travel to conventions and meetings
_____continuing education for staff
_____temporary clerical help

In many family planning agencies these items are budgeted at the top administrative level and are not the responsibility of the education staff. In other instances the educators have responsibility for the entire budget. Clear communication and a complete description of the education budget should appear in the discussions between the chief administrators and the education staff.

A discussion of income budgeting is beyond the scope of this manual. However, if you perform cost analyses for each of your objectives then you can estimate the total costs for annual budget line items in most categories by summing the costs to meet your objectives. These data will enable educators to give administrators and other budget planners in-put for

budget planning. The data will also facilitate decisions about which projects are feasible and which are not when an education staff or executive director has to choose which educational activities to pursue and which to delay until later because of limited economic resources.

A cost analysis permits rational estimates of the true cost, per participant, of offering workshops and courses. You will probably discover that you would have to charge a stiff registration fee to each participant in order to cover all of the costs for workshops and courses. But, with a cost analysis, you will know how much service you are giving away, if any, and you can present solid evidence to support requests for grants, consulting fees, and contributions.

Time management. With task analyses you can practice time management, the process of allocating the work time needed to complete tasks in order and on time. With even a rough task-analysis of each objective, you can make reasonable estimates of the amount of work which an educator can perform in a given period.

Of the 260 weekdays in a year, an educator loses about 5 days to illness, 10 to holidays, and 15 to vacations, leaving approximately 230. However, a typical 8-hour work day includes about 2 hours for coffee breaks, trips to the mailbox, personal phone calls ... leaving 1,380 hours per year (345 per quarter and 115 per month) for allocation to educational tasks.

Of course, for many family planning educators, a 5-day, 40-hour week would be a welcome relief. Perhaps these task analyses and time-management practices will provide some relief, or help increase others' awareness of just how much time a given set of responsibilities consumes. Educators covered by the Federal Wage and Hour Laws must be paid overtime or granted compensatory time; a task analysis should predict those requirements with some accuracy.

For those employees exempt from the federal wage statutes, a delicate balance must be struck between the need to get the job done and the need to prevent or alleviate burn-out (discussed in 1.3). Program administrators and boards of directors face the challenge of fulfilling ever-increasing responsibilites with the same, less, or at best, inadequately more resources. Yet burn-out leads to attrition and loss of efficiency, quality, and morale. Higher salaries for educators, or for any other family plan-

ning professional, is not a viable solution for the foreseeable future. Although occasional raises will be forthcoming, sufficient monetary incentive to motivate educators to work 50- and 60-hour weeks for months or years on end cannot be foreseen. Thus, those who influence educators' decisions about time management and resource allocation would be wise to consider careful analyses of how much can be done within a reasonable expectation of time commitments. Educators have a professional responsibility to wisely plan and efficiently use their time, and to learn to say no when their analyses support it.

Nothing in the preceding paragraphs will yield an indication of the **total amount of elapsed time** -- the number of calendar days, weeks, or months--required to finish a project. In the sample case of producing a flyer, the cost and task analyses estimated that 30-1/2 hours would complete the job. But the total elapsed time from start to finish of that project may be a month or two.

Avoiding that lag-time can pose problems, but the following methods may prove helpful. Anticipate the need to schedule meetings, planning sessions, printing jobs, and the like and alert others that you will have work for them to do. When several people must coordinate their efforts, the educator with primary responsibility can negotiate a behavioral contract with others, to determine when work will be done. A behavioral contract is an agreement in behavioral-objective form; it specifies who will produce what, the form which the product will take, and the day and time when it will be completed. For instance, in the case of the flyers I have been using as an example, the first task was drafting the text, followed by having a clerk type the draft, and then a session for reviewing and editing the typed draft. The lead educator might agree with the clerk and editor as to when the draft will be ready for typing, when it can get typed, and when the editor can meet to review the draft within a relatively short time-frame.

3.4 IMPLEMENTATION (Step 6)

Step 6: **Implement the various tasks that accomplish the program and its evaluation.**

Step 6 is the time when you will determine how easily you can accomplish steps 7 and 8, because you will actually create the data you will need to accomplish the last two steps. Accurate, well-organized records and files of the various numbers and papers which will constitute your evaluative data will greatly ease the process of gathering your data and comparing it to your objectives.

Many times while conducting an educational program I have had insights that would have made valuable additions to an evaluation report. Unfortunately, I have found that unless I commit those insights to paper I cannot recall them even a day or two later. So, I have found it helpful to record information in addition to what the evaluation of my written objectives would require, even if only in the form of short notes to myself.

When you are conducting workshops or courses, particularly when they are spread out over several weeks' time, you might find it helpful to write down some of the comments and questions which participants bring to you. This kind of data can provide great benefits when you write your evaluation report with the intent of using it during your next assessment and planning phases.

3.5 EVALUATION (Steps 7 and 8)

This is the place where all your planning, record keeping, and objective writing will pay their dividends. At this point, after completing the first six steps of the cycle, the process of assigning a value to the data which you have collected should be more straightforward than relying upon attempts to justify completely subjective opinions.

3.5.1 Step 7: **Gather and organize the necessary data and calculate any statistics needed.**

Based upon the steps you took in the planning and implementation phases, you should now be prepared to provide the data necessary to calculate any descriptive statistics you need for comparison with your own objectives and for submission of various reports to outside agencies and organizations. **Section nine: Analyzing Data** contains detailed information about the calculation of various descriptive statistics.

If you must depend upon others to provide some of the raw data or some of the totals or averages, as in the case of a family planning agency which has several clinic sites, you may want to request the data you need far enough in advance so that others can quickly provide what you need when you need it.

3.5.2. Step 8: Compare the data to the criteria established in the objectives, determine if the objectives were met or not, and report in writing to the appropriate person(s).

I strongly advocate written evaluation reports. There are two instances where I think a written report is indispensible:

1. when evaluating a major educational event, such as a workshop or a course;

2. at the end of a period of time covered by the educational program's formal written priorities, goals, and objectives.

Written reports summarizing evaluative data and interpreting its meanings are important for several reasons and uses. Writing such a report gives an educator an opportunity to step back from day-to-day duties and gain a broader perspective on successes, failures, and areas which might require more attention in the future. Public relations people can use the reports to write press releases making the public aware of what the family planning agency accomplishes in the community. Educators, executive directors, and boards of directors can use the reports for several purposes: to plan future education programs both large and small; to make decisions about allocation of resources; to help orient new staff members to the educational program; and to suggest areas which need particular attention during the next needs-assessment phase.

An evaluation report may be written in any of several different styles and formats. Minimally, an evaluation report needs to include three kinds of material. An introductory section should state information identifying the program or programs the report covers. For example, a report on a workshop should clearly identify the dates, times, places, trainers, and number of participants involved in the workshop. A report on the data collected to evaluate a year's worth of educational programs should briefly recount the process and identify the cast of characters involved in setting priorities and writing goals and objectives. It should also refer to the

minutes of committee or board meetings at which policy decisions were made and goals and objectives approved.

The second section of the report should be a straightforward comparison of the evaluative data to the behavioral objectives written in **step 2.**

The third section of the report should discuss objectives which were met and those which were not. Conclusions and suggestions related to successes and failures appropriately belong in this discussion.

Some administrators prefer that their supervisees simply report the evaluative data, leaving the process of drawing conclusions to the discretion of the administration or the board of directors. There is an equally legitimate case to be made for the idea that the people who perform the work are in the best position to understand the meaning of the evaluative data, as long as there is a clear separation in the report of hard data from discussions and conclusions.

Those who have performed the educational tasks are also in a position to report what the evaluation literature calls **unexpected effects, side effects,** or **unanticipated effects** of the programing. Almost every educational event, process, or product will cause these kinds of effects, which often reveal as much as or more than the data relating to the written goals and objectives. Frequently, interpretations of the evaluative data aimed at explaining partial or total failure to reach the objectives will yield as much valuable information about the program and the next needs assessment as the data pointing to unqualified success. A discussion of these unanticipated effects, or explanations of success or failure in terms of them, belong in the discussion section of the evaluation report. In fact, no evaluation report can really be considered complete without some of these interpretations or suggestions for future activities, programs, methods, approaches, and revisions.

Some evaluators argue that only an outside evaluator can really see the full extent of unanticipated effects or make clear and unbiased interpretations of the data gathered for evaluative purposes. The argument about outside versus in-house evaluators and their advantages and disadvantages has been ongoing for some time and is likely to continue. As is usual in such debates, there is merit and truth to the arguments offered by both factions. An outside evaluator can offer a new perspective and uncover things which the people who have been living with the program may

not see. At the same time, evaluation performed from within the agency, even by the people conducting the programs, has its benefits. Not the least of the benefits gained from having educators evaluate their own work is the psychological exercise of examining oneself.

In reality, most family planning programs must rely upon voluntary or mandated outside evaluation rather than outside evaluators hired as consultants. Thus, most family planning educators and their supervisors need to communicate clearly and discuss the evaluative process openly and regularly.

In summary, the real work in conducting the evaluation of a family planning educational effort occurs during the assessment and planning phases, at the same time when an educational program is planned and implemented. If the first six of the eight steps described above are executed as prescribed, the last two steps will follow quite easily and quickly. However, if one attempts to evaluate an education program after the fact, one loses the ability to collect much of the data which would be most revealing in the final evaluative analysis.

3.6 CONSULTANTS

There are various individuals in almost any community or state who can provide invaluable consulting assistance in the process of evaluating family planning education. Some of these people, in fact a great number, are willing to help on a voluntary basis because they can gain some benefit other than financial reward for their efforts.

For example, graduate students in various fields need practical experience to satisfy the requirements for field work, internships, or practica. Of course, any university faculty members who serve on the board of directors at a family planning agency might legitimately be approached for assistance in their areas of expertise.

Regional, statewide, and national family planning organizations employ people responsible for providing technical assistance to individual family planning agencies. Sometimes your peers at other family planning or social service agencies can provide this kind of technical assistance free of change, or in exchange for some service which they might need. Government agencies, such as health systems agencies and family planning offices, may also have personnel who can offer consultations at no expense to your agency.

Ultimately, it takes just as long, if not longer, to learn all of the various research and evaluation skills necessary to evaluate family planning education programs as it takes to become an experienced family planning or sex educator. And, just as we family planning educators call upon others when we need assistance in program development and materials preparation, we can call upon experts in other fields to make our work easier.

You will face judgments about the competency and commitment which potential consultants bring to the task. Competence is hard to assess, commitment perhaps harder. You may feel surer of both if you can negotiate a contract, if only a verbal one, in behavioral form. A behavioral contract satisfies all of the criteria for a behavioral objective (**section two**) but may require several sentences to state completely. If you doubt the competence of a possible consultant, hand the person a list of the objectives you need help in evaluating and ask, "How can we evaluate these?" The answer should immediately reveal the consultant's strengths and weaknesses.

SELF-TEST ON SECTION THREE

You will probably gain the most benefit from taking this self-test if you write your answers on a separate sheet of paper and then check them against the answers on the pages following the test after you have answered each question:

1. What is a needs assessment?

2. Rank the following methods of assessing the needs of a small population for whom you are planning an educational program such as a workshop or course from most effective to least effective in terms of how much each method is likely to reveal about the group's needs. Mark a "1" beside the most effective, a "2" beside the next-most effective, and so forth:

   _____A. Surveying the entire population through direct contact with the members of the group.
   _____B. Surveying a sample of the population.
   _____C. Interviewing the leaders of the population or colleagues who work directly with the group during an on-site visit.
   _____D. Inviting the leaders or the professionals who serve the population to your office to meet with and interview them.
   _____E. Surveying the literature to discover what others have published about this kind of educational program for this kind of population (or interviewing others who have done the kind of program with the kind of group you have in mind).
   _____F. Deciding what the group's needs are, based upon your experience as an educator.

3. Match the following concepts with respect to an educational program:

_____1.	Content	A.	The educational methodologies chosen.
_____2.	Process	B.	The facts, skills, or affective material of an education program.
_____3.	Outcome	C.	The long-term changes in a program's participants, or the extent to which the program alleviated health or social problems or concerns.
_____4.	Impact	D.	The determination of the relative efficacy of two approaches to imparting the same information.
		E.	How participants differ immediately after a program from immediately before it.

4. Should family planning educators become highly knowledgeable in research- and data-gathering techniques? Why or why not?

5. Name five libraries specializing in sexuality and family planning educational materials.

6. Define task analysis, cost analysis, and cost-effectiveness.

7. If you have total budget of $1,830.00 in expenses for a workshop to train 30 nurses in basic family planning educational and interviewing strategies, what is the cost-effectiveness of the workshop?

8. What is time management?

9. List four reasons why a family planning educator might want to commit an evaluation report to paper.

10. Name four potential sources of free or voluntary consulting help with family planning educational program planning and evaluation.

ANSWERS TO SELF-TEST ON SECTION THREE

1. A needs assessment is a process of collecting and analyzing data and other information in order to ascertain the education program needs of a community or a group at whom a particular event is aimed; a process to determine who most needs what kind of program where and when.

2. The methods are listed in descending order of effectiveness; A = 1, B = 2 ... F = 6.

3. 1 = B; 2 = A; 3 = E; 4 = C. D refers to evaluating two kinds of educational processes to impart the same content.

4. No, because such an effort would require years of study. Just as other professionals call upon us family planning educators for help with matters in our areas of expertise, we should feel free to rely upon others to share their expertise. However, we need to be sufficiently knowledgeable to ask cogent questions and to understand the complete range of tools available to us.

5. You might want to check your list against the list in 3.3.3 and add any more you happened to know about to the sixteen listed there.

6. A task analysis is a sequential list of the various steps toward accomplishing an end product with an estimation of the person-hours required of each person who will contribute one or more steps. A cost analysis is an accounting of the predicted total expenses to produce an end product in an educational endeavor. Cost-effectiveness is the ratio of unit output to total cost for a product or service.

7. $$\text{Cost-effectiveness} = \frac{\text{Total cost of project}}{\text{Numbers of units output}} = \frac{\$1,830}{30} = \$61.00$$

   in this case; which means it costs $61 to conduct the kind of training specified per nurse trained.

8. Time management is the practice of effectively using work time and efforts by planning and prioritizing the sequence of tasks on an hourly, weekly, monthly, and/or yearly basis. (Some authors and trainers use this term to refer to specific instruction in techniques to avoid time waste and increase the amount of work output per unit of time: e.g. limiting and timing phone conversations; handling each piece of paper only once; making a regularly scheduled time of the day for calls and apppointments; careful travel route planning; and so forth.)

9. a. To serve as a source of information for the next needs assessment process; b. to use as a help in orienting new employees; c. to use in on-the-job training for entry-level employees; d. to account to supervisors, boards of directors, or other agencies; e. to provide grist for the publications mill; f. to provide an opportunity to reflect upon a period of work, as a means of stepping back from the forest; g. to stockpile documentation as ammunition against attack....

10. a. Colleagues and peers; b. government agencies (US DH&HS, HSAs); c. university faculty; d. graduate students and advanced undergraduates; e. national, regional, and statewide family planning consortia and agencies; f. experts in relevant fields among the membership of a family planning program's board of directors; g. community college faculty.

section four:
LOCATING AND USING OTHER PEOPLE'S DATA

After reading this section and learning the material in it you will:

1. Be able to identify sources of data for use in needs assessments and for gauging the impact of your program on your community.

2. Have a logical method for utilizing the data available in order to make accurate estimates of facts and figures for your needs assessments and other evaluative studies.

3. Understand the state of the art in assessing the impact of an educational program in terms of vital statistics and similar data.

NOTA BENE: Throughout this section I make sample calculations which involve decimals and percentages. If you are not familiar with the use of percentages then you may want to turn to the discussion of them in 9.2.2. If you choose to perform the sample calculations along with me in this section you will probably find a small hand-held or desk-top electronic calculator very helpful.

THERE IS NO SELF-TEST FOR THIS SECTION.

Although I have tried to choose realistic examples, ALL OF THE FIGURES, PERCENTAGES, AND OTHER NUMBERS IN THIS SECTION ARE FICTITIOUS.

Strive at all times for accuracy in the figures you report, especially when you rely upon data gathered by others. Check original sources and document **everything**.

section four: LOCATING AND USING OTHER PEOPLE'S DATA

> **"Never choose** [to do a research project] **which requires you to collect data unless that is your last resort."**
>
> **--Keith E. Meredith**

This section may prove to be the most important one in this entire manual in terms of saving you time and effort. It will acquaint you with sources of information about the populations which you and your agency serve, and with figures and other data potentially useful when you want to assess needs, gauge impacts, or identify shifting patterns of social concerns. It will suggest a framework for logically applying arithmetic to the data available to arrive at useful accurate figures.

Some art is involved in the process of using data about populations or groups within a community in order to describe their characteristics, needs, and problems. Social scientists must frequently make estimates based upon estimates of other estimated guesses. Figures like birthrates, numbers of abortions, and percentages of births to single mothers which were registered to teenagers will never achieve 100 percent accuracy because they depend upon the diligence and consistent reporting of too many different people--people with diverse talents, motivations, and abilities. Even counting the numbers of births in a municipality, county, or state cannot be 100 percent accurate because some clerical errors will inevitably appear. **Therefore, be careful that your sources are accurate!**

When we notice the impact of our programs, or what we hope is the impact, we can never separate our effects from the various other social and demographic impacts upon a given set of figures. For instance, if births to single mothers drop in your county how can you be sure that your educational program had any effect? Your program must have had some effect, but how much? We will probably never know. However, we can identify shifts and trends and use our evaluative data to document claims that some of those shifts and trends were a consequence of our efforts. And, we can use our evaluative data to indentify approaches and methods which make the biggest possible impact through the wisest possible use of our limited time, money, and other resources.

As you make use of the data sources listed here, or any others you can locate, **take care to assess the accuracy and the source of the information** Seek to determine whether the information you have is the original source or selected and manipulated excerpts from the original data. Some people take a set of raw data, extract those portions which will best suit their purposes, and present it in the most favorable light possible in terms of their particular needs. Although finding the original source will require additional time and effort, you may be wise to expend that time and effort to ensure that the data you use are accurate and complete. In addition, when you examine the original source, read it critically to search for flaws in the data-collecting and reporting process. Avoid extracting tables, charts, and graphs from any source until you have read the accompanying text to search for limitations and disclaimers which may influence your use of the data.

4.1 SOURCES OF DATA

Literal mountains of data are collected annually about the US, its population, and the behavior of Americans. Here are some commonly useful sources of demographic data.

Government agencies. Federal, state, county, and municipal agencies collect data in massive amounts about the populations you serve. You can never be sure of the nature, extent, or accuracy of those figures until you wade through the bureaucracies and ask for what you seek. Sometimes you find an agency which purports to have collected exactly what you need--only to find that their figures are grossly inaccurate based upon what you know about your community. For instance, I once needed figures on the number of abortions performed in my county and wrote to the state health department for that figure. Based upon a survey of local providers I estimated that there had been about 2,500 abortions in my county the previous year, including about 850 performed for teenagers. When the state data arrived at my office they reported less than 1,400 abortions to women of all ages in my county the previous year, an obviously inaccurate and useless figure. But on other occasions I have found their data to be quite useful and seemingly accurate.

Rather than attempting to discover all possible printed sources of data, begin by writing or calling the most logical agency from the list under state, city, county, or federal government in the local phone directory. Your call may become a buck which gets passed from person to person, from office to office, but perseverance will usually pay off.

You might find the following publications useful:

CURRENT POPULATION REPORTS. Population estimates and projections. Series P-25. Population estimates for counties, incorporated places, and selected minor civil divisions (by state). Price varies. Order from: Superintendent of Documents, US Government Printing Office, Washington, DC 20402.

Every US congressional district has two designated **Repositories of US Government Documents,** literally collections of everything the US government prints other than secret materials. Call your local congressional district office and ask where the repositories are located and then seek the help of a librarian there to find the issue(s) of the above which you need.

COUNTY AND CITY DATA BOOK, 1977. US Government Printing Office, address above, 1,011 pp., $19.50. STATISTICS FOR STATES AND METROPOLITAN AREAS. Same source, 110 pp., $2.75.

The COUNTY AND CITY DATA BOOK contains 195 statistical facts about each state and 3,143 counties plus 161 items of data for 277 standard metropolitan statistical areas (SMSAs). The second citation is an abbreviated version of the larger volume. These should also be available from your local Repository of US Government Documents. Find one and photocopy the few pages you will need rather than buying the document.

FAMILY PLANNING SERVICE STATISTICS ANNUAL REPORT, US Agency for International Development, Office of Population, Washington, DC.

These volumes are issued annually and began with the 1974 edition. They use clinic reporting systems to categorize international family planning programs by contraceptive users and program activities.

DIRECTORY OF FEDERAL STATISTICS FOR LOCAL AREAS: A Guide to Sources. US Government Printing Office, address above, 363 pp., $5.50.

The Directory gives table-by-table descriptions of subjects and reference sources from 350 reports issued by 36 different federal agencies.

STATISTISICAL ABSTRACT OF THE UNITED STATES. US Department of Commerce, Bureau of the Census. Order from US Government Printing Office, address above.

This annually issued volume, in its 100th edition in 1979, contains a wealth of information about the US, from the most common to the most obscure.

WORLD ALMANAC AND BOOK OF FACTS. Newspaper Enterprises Association, New York. Printed and distributed throughout the US through newspapers and on paperback book stands.

DEMOGRAPHIC, SOCIAL, AND ECONOMIC PROFILE OF THE STATES: Spring 1976. Series P-20 of Current Population Reports, cited earlier.

The business sector. Banks, savings and loans, real estate groups, chambers of commerce, and businesspersons' groups, as well as organizations of all of them, issue reports which include local demographic data, income levels, racial and ethnic breakdowns, and other population characteristics. These volumes are often free or sold for a nominal fee. Although they are not written for our purposes, the data which these volumes contain have proven useful to me on many occasions.

Marketing researchers and pollsters. Both the print and broadcast media base their advertising rates and their programming decisions on data gathered about audience ratings, subscription and circulation rates, and the characteristics of their audiences (age, income, educational level...). These data might help you make decisions about where to spend your advertising and public service announcement (PSA) resources. An acquaintance at a local radio or TV station can probably share these data with you.

Universities and colleges. Institutions of higher education and their consortia compile and publish many kinds of reports on economic, educational, and demographic data as a service to the community or to fulfill their commitments to research and development. Their libraries' reference departments should have ready access to copies of their work.

Librarians. In fact, libraries in general contain amazing collections of materials which could include the figures or facts you need. As a rule of thumb, do not assume that what you seek does not exist until you have contacted every library in your area and asked for it. I have been amazed more than once when a librarian handed me something which I was sure could not possibly exist.

4.2 USING THESE DATA

Rarely will you encounter the exact figure or fact you require in a report or an article. However, if you can locate enough related facts and figures you can often analyze and logically combine them to produce an accurate figure.

Consider the following four-step process for using what you have to get the fact or figure that your work requires.

1. **What is the fact or figure you want?** Attempt to specify as exactly, as behaviorally, as possible the quantity you are seeking.
2. **What facts or figures do you have which relate to question 1?** Collect as many figures as you can related to what you seek to verify or derive.
3. **How can I combine what I have to get what I seek?** Think in terms of, "If..., then...." statements. For example, if there were 600 births to teenage mothers in my county last year and 30,000 teenagers living here, then there were 20 births for every 1,000 teenagers. (600 births/30 thousands of teens = 600/30 = 20 births/1,000 teens.)
4. **If you cannot combine what you have to get what you seek, then what other figures would you need to complete the logical chain?** Suppose that a new car has a speedometer marked in kilometers per hour (KPH), but no miles per hour (MPH) markings. How many KPH is 55 MPH? How many K are in an M? The almanac says that 1 K = .675M. Therefore,

$$55 \frac{M}{hour} \times \frac{1K}{.675\ M} = \frac{55}{.675} \frac{M \times K}{Hr.} = 81.5 \frac{M \times K}{Hr. \times M} = 81.5\ K/Hr.$$

Notice that the M in the numerator of the fraction cancelled with the M in the denominator, leaving me with KPH. If I had multiplied by .675 instead of dividing, my units would have been M X M/Hr. X K and I would know that I had made a mistake.

Example 1. Suppose that you want to know the birthrate for black teenagers in the county. I know that : (1) 1,000 babies were born to black teenage mothers in this county last year; (2) 22% of the population in my county is black; (3) the county population is 512,205. Birthrates are expressed as the number of births per 1,000 population. Therefore, if you know the total black population in the county, you will be getting a good start. And, 22% of 512,205 is 512,205 X .22 = 112,685, the number of black people living in the county. But that figure represents all black people in the county, not black teenagers. Therefore, what portion of the black population is teenagers? A local demographic breakdown published by a bank says that 26% of the blacks in the community are between the ages of 12 and 20. Therefore, 26% of 112,685 blacks are teens; 112,685 X .26 = 29,298.

Finally,

$$\frac{\text{1,000 births to black teens}}{\text{29.298 thousands of black teens}} =$$

34.1 births per 1,000 black teens in the county. Notice 29.298, not 29,298, in the denominator of the final fraction; the units for the answer, births per 1,000 population, express the denominator in units of 1,000. (Calculating percentages--see 9.2.2.)

Example 2. How many unmarried teenagers in this state are exposed to unplanned pregnancy because they are sexually active and do not use any reliable form of contraception? The first step would have to be finding out how many unmarried teenagers live in the state. Suppose the following facts:

1. 10 million people live in the state;

2. 10.2% of the people in the state are aged 12-15;

3. 9.9% are aged 15-19;

4. Among teenagers in the state, 12% are married.

Facts 2 and 3 show that 10.2% + 9.9.% = 20.1% of the states's residents are teenagers: 10 million X .201 = 10,000,000 X .201 = 2,010,000 teenagers living in the state. Since 12% of the teenagers are married, 88% are not (taking the data literally, leaving out divorced and widowed people). Therefore, 88% of 2,010,000 teenagers are unmarried: 2,010,000 X 88 = 1,768,800.

Suppose an article which appeared in a nationwide publication has said, "Nationwide, 54% of unmarried teenagers are at risk of unintended pregnancy each year because they have sexual intercourse without using any form of contraception." There are two problems: Does the population of the state match the nationwide population used for the article? And, does using "no form of contraception" match the original question about "no reliable form of contraception." Obviously, you change the original question, or assume that an unreliable form of contraception is little better than no form of contraception. One would have to read the entire article in that nationwide publication, however, and compare the population of the state with the nation's population in terms of those things (race, ethnic origin, religion ...) which the authors of the article seem to think affect

the risk of unplanned pregnancy from unprotected intercourse. For instance, if the authors of the article found that whites are at higher risk than blacks nationwide, you would want to make sure that the state's population matched the nation's in terms of racial breakdowns. If things matched, their 54% of 1,768,800 is 955,152, which could be reported as "about 1 million" or "just below 1 million" or "nearly 1 million" unmarried teenagers at risk of unintended pregnancy in the state.

If the state's population did not match the nationwide sample in terms of some significant characteristics, you would have to divide 1,768,800 unmarried teenagers into the relevant groups, apply the correct percentages of risk from the article and then add the results to get a total figure.

Example 3. How many teenage pregnancies did this family planning program prevent during the past year? Assume that:

1. The family planning program serviced 1,235 teenage patients last year;
2. 82% of those patients use the birth control pill, 8% condoms and foam, 5% use diaphragms, and the remaining 5% have IUDs;
3. One can ascertain the use-effectiveness for the general family planning population for each of those methods of birth control.

You could proceed by multiplying the 1,235 (teen patients) by the percentages of patients who use each method and then multiply those figures by the use-effectiveness to get numbers of teenagers who avoided pregnancy using each method. The total of those figures would yield the number of teenages who avoided pregnancy through the program's efforts.

BUT, from reading the literature on family planning, you know that teenagers have a strong tendency to quit using their methods of birth control, to stop taking their pills or to take chances with barrier methods more frequently than the general family planning population. Consequently, using the method outlined in the last paragraph will yield inflated figures. In addition, condoms and foam are popular methods of birth control among teenagers, used alone or in combination. How many teenagers heard one of the program's educators give a talk, then began to buy condoms and foam at a drug store without ever coming near the clinic?

Carefully documented use-effectiveness figures for the pill, the IUD, the diaphragm, condoms, and foam **for teenage populations** could be used to apply the method suggested in the first paragraph of this example. Then

you could document **at least** that many pregnancies averted by the program. However, a search of the literature on family planning and teenage pregnancy fails to reveal any research on the use-effectiveness of various methods when used by teenage patients. Consequently, you cannot document the answer to the question until someone does some more research. Therefore, you would have to say that you cannot accurately answer the question.

4.3 THE QUESTION OF IMPACT

Example 4. Suppose that your mail today includes a new set of tables from the state health department, division of vital statistics, showing the numbers of births to single mothers in the state, county by county. The state as a whole saw a 10% increase in births to single teenage mothers for 1980 compared to 1979. Your county, however, had a 2% **decrease** in the same figures. Can you claim that your education program caused that decrease?

Before you could make such a claim you would have to carefully consider various factors which could account for the decrease in your county. What are some of the factors which could change the numbers of births to single teenage mothers? Since the figures from the division of vital statistics are expressed as numbers of births, if the number of teenagers in the county or state increased or decreased then that could explain the change. Therefore, you would want to convert the figures for numbers of births into birthrates, expressed as numbers of births per 1,000 single teenagers in the population of the county and the state.

Suppose your converted figures to birthrates still showed about a 10% increase in the birthrate statewide and a 2% decrease in your county. You have eliminated the factor of increased or decreased numbers of teenagers. What else could affect the figures? Fewer teenagers are having sex in your county than in the rest of the state? A possible but improbable explanation, which I would dismiss. Therefore, if the rate of births to single teenage mothers is decreasing in your county while it increases in the rest of your state, one of the following, or a combination of them, must explain the figures:

1. More teenagers in your county get married between conception and birth than in the rest of the state, thus resulting in those births being registered as births to married women.

2. More teenagers who get pregnant while living in your county move to another location to have their babies than for unmarried teenagers in the rest of the state.

3. More teenagers in your county have abortions than in the rest of the state, thus representing no real drop in pregnancies but appearing as a drop in the figures for births to single mothers.

4. More single pregnant teens in your county have miscarriages than in the rest of the state.

5. More single, sexually active teenagers use birth control in your county than in the rest of the state, a trend which has developed over the past year.

In the absence of extenuating circumstances you could dismiss numbers 2 and 4 as highly improbable. I would seek to verify the possible vailidity of number 1 by consulting the statistics for marriages in the county, to determine if there was a sudden increase in the numbers of marriages registered for teenage women. I would similarly seek data on the numbers of abortions performed for teenagers in the county, to compare the increase or decrease from 1979 to 1980.

Assuming that all of these factors proved insufficient explanations of the decrease in births to single mothers, you could conclude that number 5 on the list must be at least a partial explanation. Then, seek to discover if other education programs, such as a new family life or sex education program in the local schools or some other teenage-serving agency, could have had an impact. But at that point you have clear evidence of an impact from some source.

I would never claim that my program was responsible for all of the decrease, or lack of increase. Too many other factors might feed into the complex system which determines the number of teenage pregnancies to draw a hard and fast conclusion. Just as certainly, however, some of the visible impact of education programs must have been a result of your program's efforts.

Thus, to attempt to explain observable impact, you logically examine all possible explanations for the data you have and eliminate them or attribute some of the impact to them. This "back door" approach is the weakest kind of attempt at documenting impact from your education program, but until we have much more time, money, and better techniques for measuring societal characteristics, it is the best we have. As educators, we are a small part of the totality of influences, needs, stimuli, motivations, peer pressures, attitudes, and self-images which influence people's decisions

about how they behave. That does not mean to say that educators do not or cannot have a profound impact upon an individual's behavior or life. It simply means that we need much better methods of measuring and documenting the impact before we can quantify it, map its path, and maximize its effectiveness.

Of course, **if you have an isolated group of participants** in an education program, and you follow them for a period of time after they attend your program, you are not using the "back door" approach. Rather, you are doing longitudinal surveys (see 7.1) or experimental research **(section 8)** to measure impact.

Example 5. Suppose, for instance, that you have just conducted a series of workshops for parents on how to improve communication about sexuality between the parents and their children. You might guage the outcome (see 3.1, p. 36) of the program through a questionaire on through interviews with the parents, designed to measure the effects of the training upon their communications with their children about sexuality. You might also administer a knowledge or attitude test to guage the outcome in those terms.

To guage the impact (sse 3.1) you might keep in regular contact with the families who attended your training sessions for two, three, or even five years. During that time you would want to guage their communications patterns again, as you did with the outcome evaluation. You could also compare the numbers of pregnancies experienced by the children whose parents attended the training; the number of cases of sexually transmissible diseases among the children; the numbers of these young people who choose not to become sexually active; or whatever else indicates a valid measure of your training program's impact.

For instance, with an identifiable and isolated group, the impact of the program can be guaged much more specifically. As always, the accuracy of the evaluation data collected to measure impact depends heavily upon the reliability and validity of the instruments and techniques used to gather data. See 5.3 and 6.2 for discussions of validity and reliability.

section five:
HOW TO WRITE QUESTIONNAIRES

After learning the material contained in this section you will be able to:

1. Write questionnaires based upon ten guidelines which will improve the quality of the instrument and increase the probability that you will gather valid, reliable, and usable data.

2. Identify and write samples of the following item types and state their uses, advantages, and disadvantages:

   a. Likert-type
   b. rating scales
   c. semantic differential
   d. cloze
   e. open-ended
   f. listing
   g. check lists
   h. Q-sorts

3. State the reasons why pretesting a questionnaire is important and identify those instances in which pretesting is indispensable.

4. Define the importance of validity and reliability as they relate to questionnaire instruments.

5. Interpret the meaning of the data gathered by various questionnaire items.

If you are already familiar with this material you may want to turn to **section six.** If you are not sure, or if it has been a while, you may want to turn to the self-test.

NOTA BENE: Some of the concepts contained in this section and in section six are closely related. You may find it helpful to learn or review the material in section six before you apply the principles in this section to a specific project. In particular, the discussions of validity and reliability in this section are brief; those in section six more detailed.

section five: HOW TO WRITE QUESTIONNAIRES

"What do you mean nobody will fill out this questionnaire? It's only forty-two pages!"

--A College Professor

Questionnaires can be one of the most efficient and effective ways to gather evaluative data either for needs assessments or for evaluating participants' perceptions about an educational program's process, outcome, or impact. However, in order to ensure that your questionnaire data will be acceptable as valid evidence from which you can draw conclusions, you will need to take care when you prepare it. All too often, professionals simply type up a few questions, make some photocopies, and distribute them to a group of people from whom they want information.

There are professionals who devote much of their working lives to the fine art and science of constructing questionnaires. However, they follow a relatively small number of basic principles, which you can learn in a reasonably short time.

The first step in writing the questionnaire is to decide exactly what information you want the questionnaire to gather from the respondents. You might begin this process by making a list, in behavioral-objective form, of the things you need to know about or from the respondents. You also need to carefully consider what kinds of demographic information you will need about each respondent: age, sex, race, religion, occupation, employer or school, educational level, and the like.

Next, you need to translate your list of informational needs into specific questions, or, as they are commonly called, **items**. A wide variety of item types are at your disposal. Section 5.2 will give you the complete list along with discussions of the advantages, disadvantages, uses, and abuses of each item type.

5.1 GUIDELINES FOR CONSTRUCTING QUESTIONNAIRES

Before describing the various item types, however, I want to suggest some basic guidelines and principles for you to follow as you write questionnaires.

Guideline one. Write as few items as possible. Questionnaires are subject to the same validity and reliability considerations as tests (see

6.2). A long questionnaire is less reliable than a shorter one which gathers the same information because respondents get bored quickly and begin to write down the first answer which comes to mind, or to skip items entirely. The length of a questionnaire is less of a problem when you administer it during a face-to-face or telphone interview, provided that you alert the respondent to the fact that you will need several minutes of his or her time.

Sometimes you can take what seems like a lengthy questionnaire and produce it in a shorter format by typing the original on legal-size paper and then having the original reduced on a photocopy machine from legal-size to letter-size paper. Obviously, this will not reduce the number of questions, but it will give the respondents the impression that the questionnaire is not quite so long.

Guideline two. Ask only one question per item. Sometimes questionnaires with items look like this one:

> Do you agree or disagree with the statement that the facilitators for this workshop were knowledgeable in the subject matter which they covered and that they encouraged questions?

That item is two questions: one question about knowledgeability and one about encouraging questions. Not only will respondents have difficulty in answering the question, but also, if they do answer it, you will not know which of the two questions they have answered.

It is legitimate to ask two questions in one item if the item asks for a narrative response:

> In the space below, please tell us two things about this workshop: (a) was there anything which we did not cover that you would have liked us to cover? (b) do you believe that what you have learned during this workshop will be useful to you when you return to your job?

However, even a question such as this one might be better divided into two items.

Guideline three. Pay close attention to diction, syntax, grammar, and vocabulary. Poor grammar will distract the respondents' attention from the intent of the question. Respondents will often "grade" the item instead of answering it.

Questions which sound awkward, or which contain professional jargon, will not gather reliable responses. Here is an example of what I mean:

> Tell us in your own language how well you did not like or did like this past two days of workshop and why you rated us good or not in your own opinion and words.

Or, even worse:

> What are three things which a pg. c. should tell a nullip o.c. pt. during a methods rap?

Such jargon and poor grammar will gather unintelligible results. For this and other reasons, it is important to have other people read drafts of questionnaires before having them typed and duplicated.

Sometimes questionnaire writers decide that they need to ask their questions in the most formal language and sophisticated vocabulary possible:

> Following circumspect cognition of your rejoinder, insinuating nothing, please report the endpoint of your thoughts on the following statements (kindly eschew congnizance of affect).

That admittedly ridiculous example makes the case for writing questions in the simplest language possible. This can be a crucial consideration when administering questionnaires to lay people because the average adult in the United States reads at about the sixth-grade or seventh-grade level. That means that the average adult has difficulty reading the evening newspaper.

Guideline four. State what you want to know in simple declarative or interrogative sentences, avoiding compound, complex, verbose items.

> BAD: Feel free to use the back of this paper or to attach additional sheets if necessary to give us your reactions to this workshop or its content notwithstanding the questions you have already answered or any marginal comments you found it necessary to make beside the questions above.

> BETTER: Use the back of this sheet to make any comments you want to make about the workshop. You may attach more sheets if you need more space.

> BAD: If you are willing, please share with us any feelings you have about us, about our workshop, about what the workshop covered, because we really think it's important to get your opinion on such things.

> BETTER: Please use the back of this sheet to tell us anything you would like us to know about the workshop.

Guideline five. Use a mixture of items in each questionnaire you write. Mixing item types in a questionnaire improves the reliability of the instrument by reducing the inevitable boredom respondents feel when answering a whole series of similar items. It also prevents **response sets.** When respondents receive a list of items to which they are asked to express opinions like "agree, disagree, strongly disagree," and so forth, they sometimes simply circle one response to a whole list of questions without even reading the items: a response set. Mixing item types also prevents the **"indecisive funnel."** Sometimes a respondent, when faced with a long list of agree-or-disagree items, will begin with strongly decisive opinions at the top of the list, and then become less decisive toward the end of the list, tending to mark more and more items with less and less decisive responses. When you examine such a questionnaire, the responses which the subject has marked are spread along the ends of the continuum near the top of the page and form a V- or funnel-shaped grouping near the bottom of the list. Studies have shown that this occurs because people dislike to appear opinionated, to have too many decisive attitudes at one time.

Guideline six. Be polite and generous. Always begin the questionnaire with a short, clear statement which asks the respondents for their assistance and explains the reasons why you want their answers to your items.

Give ample space to write answers to questions requiring narrative responses. If you give respondents no more than two or three blank lines to write a paragraph then you either give them the impression that you don't really care about their opinion, or you frustrate them when they try to adequately answer your item.

Guideline seven. Avoid assumptions, presumptions, and judgments. As you write your items consider the wide range of experiences, attitudes, feelings and knowledge which your respondents will bring to the task of completing your questionnaire. Following this guideline will increase your chances of getting valid and reliable responses, and decrease the number of respondents who omit answers to some of your items. When writing questionnaires you need to be aware of your own values and biases, and take care not to include them in your questions. For example, give either all of the possible responses to a question or none:

BAD: Do you think teenagers should have their parents' or guardian's permission before obtaining contraceptives?

BETTER: Do you think teenagers should have to have parental permission to get contraceptives or do you think teenagers should be able to get contraceptives without their parents' knowledge or consent?

Similarly, do not assume that you know how people feel about a particular issue or event. If you include items which ask for positive opinions, include a separate item which asks for negative opinions.

When you write questionnaires, it is easy to presume knowledge or experience on the part of the respondents which they may or may not have. If you make such presumptions you may divert the respondents from the task of answering your items by injuring their egos. An item such as this one presumes that the respondents have a specific kind of information:

BAD: Where can teenagers go in our community to get birth control services?

BETTER: Do you happen to be familiar with the agencies which offer family planning services for teenagers in this community?

The latter method of phrasing that item calls for a yes-or-no response. Thus, if you want the respondent to list the names of family planning agencies, then you might word the item in this fashion:

Please list below the names of any agencies you happen to know of where teenagers can get birth control services.

Some items imply or allow inference of judgments about a particular kind of behavior. For example, suppose that for some reason you want to know if your new teenage patients have had sexual intercourse when they arrive at your clinic. Each of the following items implies, or permits inference, of judgments about sexual experience.

Are you a virgin?
How long ago did you start having sexual intercourse?

A better way to get the same information might be to ask the respondents:

Have you ever had sexual intercourse?

Of course, that item assumes that the respondent is familiar with the term "sexual intercourse." Substituting terms such as "making love" can confuse your data because the meaning varies from person to person. You could, of course, add to the last sample item a sentence defining what you mean by sexual intercourse.

When conducting a needs assessment among a population of potential participants in an educational program, avoid implying judgments about knowledge or skills which they may or may not already have. For example, I have used the following kind of item on several needs assessment questionnaires:

Topics	I need a refresher	I have this skill already	This is not related to my job
Empathic listening	_____	_____	_____
Basic interviewing skills	_____	_____	_____
.	_____	_____	_____
.	_____	_____	_____
.	_____	_____	_____

When I first started using items like the last one, I was concerned that I would not find out what the potential participants desperately needed to know. However, I found that many of the respondents would write notes next to a particular topic saying things like: "I need a lot of this," or "Definitely cover this."

Also, try to begin the questionnaire with closed items, that is, items which require very little writing or which call for simply making a check mark or circling a response. If you save items which ask respondents to write sentences or paragraphs until later in the questionnaire then you are likely to get responses to your closed items.

Miller (1977) suggested that items which ask for positive and then negative opinions be listed with the positive request first. Suppose you want to ask respondents what they liked and what they did not like about a workshop. Miller's suggestion is that you ask them what they liked first.

Guideline nine. Begin construction of your questionnaire as early as possible. There are several good reasons for this, the most important of which are covered in 5.3. In addition, you can increase the number of questionnaires which are returned as well as the number of items completed on each questionnaire by printing the questionnaire so that all of the copies are clear and easy to read. If you have a lengthy questionnaire, consider printing it in the form of a booklet with an attractive cover and a title which will provoke the respondents' interest.

You may also need sample copies of the questionnaire for others to review. Most groups and organizations, particularly institutions such as churches and schools, will want time to review your questionnaire before you distribute it. If you have volunteer consultants who are also university students, the university may have requirements that the faculty members or committees review the questionnaire before it is distributed, particularly if the questionnaire pertains to a thesis or to dissertation research. Recent federal regulations require that any research conducted using human subjects have prior approval of a universitywide committee to ensure protection of the subjects' rights and interests.

Guideline ten. Treat questions of confidentiality, privacy, and volunteerism ethically, clearly, and straightforwardly. Unless some compelling reason requires that you ask respondents for their names, do not. If you need to be able to match questionnaires, such as pairs completed by a group both before and after a program, have respondents use easy-to-remember but anonymous codes (e.g. the last four digits of a credit card, checking account, or Social Security number; the month and day of some ancestor's birth; the last four digits of a phone number they call often).

The introduction to the instrument should clearly state who will have access to the questionnaires and should be a strictly truthful statement treated as an ethically binding contract between you and the respondents. Not only does this practice protect the respondents' privacy and other rights, but also improves reliability and reduces the number of omitted items on the questionnaires.

Obviously, any person has the right to refuse to answer any item on a questionnaire, whether they are responding in writing or orally during an interview. Evaluators and researchers commonly assume that potential respondents are aware of this right and act accordingly. But in the sensitive areas often covered by questionnaires related to human sexuality, the strictest ethics would dictate a statement of that right on the instrument, particularly when administered to relatively naive or inexperienced people.

Questionnaires should never ask respondents for intensely private or intimate revelations, such as sexual-history information, unless some strongly compelling reason necessitates it. In cases when such facts are

requested, every possible safeguard of privacy and confidentiality should be instituted. Aviod asking respondents to reveal their identities--or any information which could conceivably lead to their identities--on such questionnaires.

Requesting intimate details of people's lives can lead to other problems. Individuals, groups, or whole communities have reacted strongly and loudly to items asking for private, sexual information from respondents, particularly when administered to minors. Besides that, the reliability of such items is questionable at best, no matter what steps have been taken. Before undertaking to construct or administer any questionniare which is potentially controversial in nature or possibly embarrassing to respondents, enlist the aid of carefully selected and fully qualified consultants. Also, carefully discuss with your administration the need to involve board members in the process.

5.2 ITEM TYPES

The following brief catalogue of item types contains those kinds of items that you might find useful in constructing questionnaires for your evaluative data collections. In some instances, such as when you are conducting complicated research projects or constructing instruments such as an attitude inventory, you may want to refer to textbooks on those subjects, or to enlist voluntary or paid consultants.

If you decide to consult some of the references listed in the bibliography, you may need to be aware of the fact that social scientists have not yet agreed upon a common language to describe what they do. Although writers in different fields may be dealing with the same ideas, they sometimes use seemingly unrelated terms to refer to the same material or the same terms to refer to different material from one book to another.

Each possible item type has its advantages and disadvantages. Since different purposes require different types of items, this catalogue is not arranged in any particular order.

Likert-type. Almost every questionnaire contains one or more Likert-type items. They consist of a **stem** (question or statement) and two to five response choices, usually in the form of "agree or disagree" and variations of those options:

1. This workshop met my expections.

   strongly agree agree disagree strongly disagree

2. The facilitators knew their subject matter.

strongly agree agree disagree strongly disagree

You can use any of several kinds of continuums other than agree/disagree: important/unimportant, always/never, children/adults...:

Please indicate how important you think it is for this course to include lessons on each of the following topics by circling the response which most nearly expresses your opinion:

Skills related to budgeting family income and expenses.

very important important unimportant very unimportant

In order to improve our clinical and educational services, we need to know how often you get the kind of services listed below. Please circle the answer which best describes your experience here.

During the interview before my exam, I have time to ask questions.

always frequently sometimes rarely never

There is a difference of opinion among experts as to whether or not a Likert-type item should contain a neutral response category: no opinion, I'm not sure, neither agree nor disagree, neither important nor unimportant. Those in favor of including a neutral response category say that it decreases the number of omitted items because respondents do not have to express a long series of definite opinions. Those who oppose a neutral response category claim that it creates a haven for indecisive respondents and leads to response sets and the "funnel effect" (see 5.1).

Response sets and funnel effects are almost inevitable if you decide to include a long list of Likert-type items in a questionnaire. There are some things that you can do to help reduce or detect these kinds of problems. You can alternate the order in which you list the response categories after the questions. (For odd-numbered items you begin the list of response categories with "strongly agree" and for even numbered items you begin with "strongly disagree." You can also begin several questions in succession with "strongly agree" and then suddenly shift for a question or two and begin with "strongly disagree.") The technique of alternating response categories may only slow down the process of circling one kind of response category for some participants, who will notice the switching and

simply circle opposite ends for alternating questions. The sudden reversal technique will help you to indentify those questionnaires which contain response sets but will also leave you with the problem of deciding how validly the responses on that paper reflect the opinion of the respondent.

In order to check the reliability of the responses, some professionals include pairs of questions which state the opposite of each other in a long list of Likert-type items. For example, a list of agree/disagree Likert-type items might contain the statement, "The facilitators were very knowledgeable," and, much further down the list, "The facilitators did not seem to know very much." Clearly, if a respondent is expressing the same opinion and conscientiously considering each item, then the respondents who agree with the first statement would disagree with the second, and vice versa. This technique requires that pairs of statements be worded in language which is not so similar that the technique is too obvious but with words and phrases that connote opposite meanings. Some people literally devote years to the process of writing these items, pairing them accurately and analyzing their correlations (see 9.2) in order to ensure that they can make valid and reliable interpretations of their results. Consequently, given the usual limitations upon a family planning educator's time and resources, avoid long lists of Likert-type items.

Even a short list of Likert-type items can consume a fair amount of space on a questionnaire. In order to shorten the space needed, some people abbreviate the response categories with letters (ABCDE) or with numbers:

> Please indicate the degree to which you agree with the following list of statements. If you:
>
> strongly agree circle 5;
> agree circle 4;
> are not sure circle 3;
> disagree circle 2;
> strongly disagree circle 1.
>
> 1. This workshop will help me on my job. 1 2 3 4 5
> 2. I would recommend this workshop to others. 1 2 3 4 5

Sometimes the numbers are used for statistical analysis of the data collected by the questionnaires, since numbers make it easier to calculate

averages and the like. However, research studies on questionnaires have shown that respondents attach different meanings to letters and numbers than to the response categories which they represent. You can still use the abbreviated form to shorten the amount of space which your Likert-type questions will require if you simply abbreviate the response categories: strongly agree becomes SA, and so forth. This procedure does not reduce the validity of the results and still allows you to assign numbers to each response category later, to assist in any statistical analysis.

Rating scales. When you want to ask respondents for finer gradients of opinion than you can get with Likert-type items, a rating scale may suit your purposes. A rating scale may be viewed as a Likert-type item to allow more categories of response, usually six to eight, but sometimes as many as ten or twelve.

The term "rating scale" sometimes applies to complicated social-scientific research techniques. Some texts describe rating scales which are actually lengthy questionnaires in themselves (e.g., see Miller, 1977). In case you are searching in other sources for information about constructing or using lengthy rating scales, Likert, Thurstone, and Guttman worked in that area.

If someone says, "On a scale from one to ten, where would you rate that movie?" that person has constructed a rough form of rating scale. Of course, in order for the respondent and the questioner to mutually understand the quality of a "six" rating, they would need to agree upon some criteria for rating movies: acting, editing, cinematography, plot, costumes, and so forth. In order to simplify their rating process, the two might conceive of a ten-point scale for each of those qualities. Then they could rate a movie on each of those qualities and arrive at a final rating by summing or averaging their ratings on each quality. An instrument consisting of several items designed to rate something based upon a composite of important components is the type of rating scale described in the preceding paragraph. The individual items on that instrument are the kind of items I want to discuss here.

To construct a rating-scale item, begin by defining as narrowly as possible the concept you want to rate. Then, try to conceive of the two extremes of that concept. Next, construct two or three gradients between the extremes. For example, suppose that you want to evaluate the useful-

ness of the skills that participants learned during a workshop on basic interviewing. If you wrote a questionnaire item which said, "When you return to your job, how often will you use the skills you have learned here?" then you would get a wide variety of answers which would be hard to summarize and impossible to average. Either of the following two items would get the data you need and give you more options when you analyze your responses:

> How often will you use the skills you have learned during this workshop when you return to your job? Please circle the number on the scale below which best approximates your opinion:
>
> 9 8 7 6 5 4 3 2 1
> Hourly Daily Weekly Monthly Never
>
> (With the same stem and instructions to the respondents)
>
> 7 6 5 4 3 2 1
> Daily Weekly Monthly Never

Either of these two items would yield more definitive data than a Likert-type item which gives respondents choices such as "always, frequently, sometimes, never."

Rating-scale items give you greater precision but less reliability. Suppose you give the same person the same questionnaire with a rating-scale item in it on two occasions several days or weeks apart. That person is less likely to mark the same response on a rating-scale item than to mark the same response on a Likert-type item.

Rating-scale items require that respondents concentrate upon their answers. As a result, if you put more than two or three rating-scale items together, respondents become careless in considering their responses and their reliability drops even further.

You may find rating-scale items most useful when dealing with questions to which there are a wide variety of answers. For example, suppose you are doing a speaking presentation about the right to choose safe, legal abortion. You want to gauge the impact which your presentation has upon the members of the audience. One of the impacts you want to have is to convince the audience that a decision about an abortion should be left to the pregnant woman and her physician. A rating-scale item such as this one, administered both before and after your talk, would reveal shifts in opinion during your talk:

How late in a pregnancy do you believe that a decision about aborting a pregnancy should be left solely to the woman and her doctor? Please circle the number of the month on the scale below which you think is the LAST month during which the decision should be left to the woman and her doctor. Circle 0 if you think the decision should never be left to the woman and her doctor. Circle 9 if you think it should always be left to the woman and her doctor.

0	1	2	3	4	5	6	7	8	9
Never									Always

You could amplify this item by labeling points along the scale, such as when fetal heartbeat starts, when organs have formed, when quickening occurs, and so forth. This rating-scale item will obviously not reveal the reasons behind any individual's choice. Therefore, you might want to follow the above rating scale with an item which asks about the reasons behind each person's response. Then, you would have data to help you choose the content of your talk.

Semantic differentials. Omitting intermediate gradients of a concept from a rating or a Likert-type item creates a semantic differential. Respondents are asked to choose between antonyms to describe something.:

In each of the pairs of descriptors below, please circle the one which you would be most likely to use when you describe the workshop you just completed:

useful	useless
interesting	boring
organized	disorganized
purple	pizza
fun	work

I used that item on an evaluation questionnaire and literally three-fourths of the people who completed the questionnaire responded that the workshop was more purple than pizzalike in the fourth pair! Inserting such an obviously ridiculous response in a list of semantic-differential pairs creates what is called a **lie item.** Lie items of almost any type may be inserted into a questionnaire for the purpose of determining whether or not respondents are reading the items or simply marking answers.

Semantic differentials like the example above are almost useless in evaluation quesitonnaires except when you want a lie item. If your lie item reveals that large numbers of respondents are simply marking answers, then you know that the results of your questionnaire are very unlikely to

item reveals that large numbers of respondents are simply marking answers, then you know that the results of your questionnaire are very unlikely to be reliable. Semantic differentials have their appropriate uses, particularly in psychometric instruments such as personality inventories.

Cloze items. A declarative statement with one or more blank spaces for respondents to fill is a cloze item:

> When I arrived for this workshop I thought I would be _____ during the sessions. Now that the workshop is over I'd say that I was _____.

I would rarely use cloze items unless I wanted the respondents to think about something as a preliminary process to answering items later in the questionnaire. For example, you might decide to put one or two cloze items immediately preceding an open-ended item asking for a written narrative response.

The same holds true for **sentence completion** items, items which begin a statement and ask a respondent to complete it:

> Please complete these sentences:
> I am different from when this workshop began because...
> During this workshop I learned...
> After your talk today I know more about...

Cloze and sentence completion items may have educational value other than for the purpose of collecting evaluative data. They do provoke participants to think, perhaps solidifying and reinforcing thinking processes and attitudinal changes which occurred during the educational program.

Open-ended. Asking your respondents to write a narrative in the form of an invitation to comment on something gives you very little in terms of results which you can quantify. However, open-ended items serve two important purposes. First you will often get very revealing information from an open-ended question--information which would be difficult or impossible to infer from the collective responses to a group of Likert-type or rating-scale items. Second, they are a way to convey the message to the respondent that his or her opinion matters to you. For both reasons, I always include some open-ended items in every assessment questionnaire which I distribute.

Avoid writing paragraph-long open-ended questions unless you need some space to explain your thoughts. If there are several things which you

would like your respondents to tell you, then divide those concerns into separate items.

Listings. This variation on the open-ended item type asks the respondent to list a specific number of sentences or phrases:

> Please list three topics which you would very much like to see included in this workshop. If you met a friend when you leave this workshop today, how would you describe your experience? Please list three adjectives or descriptive phrases you would use:

Check lists. Asking your respondents to complete a check list can be useful in several instances. Check lists permit more than one response to a particular question. You can also use a check list to avoid the pitfalls of semantic differentials, arranging them in columns in random order and then asking your respondents to check as many as apply.

If you have a list of Likert-type items to which you want to offer two or three possible responses, then a check list may serve your purposes better:

> The sections which I currently envision including in the evaluation manual are listed below. Please place a ✓-mark beside those sections which you think would be of value to you, an X in front of the sections which you think would be of little or no value to you, and leave blank those about which you are not sure.
>
> ✓-mark = valuable; x = no value; _____(blank) = not sure
>
> _____1. What is evaluation and why do it
>
> _____2. Setting priorities, writing goals and objectives
>
> _____3. How to write questionnaires
>
> ...

Q-sorts. This procedure is not really an item but a list or collection of statements or phrases which you ask someone to sort. Each respondent is given a list of items, preferably typed or printed on index cards. Then each person is instructed to sort the items into upper and lower halves in terms of some construct, like importance/unimportance, liberal/conservative, now/later, or the like. Next, respondents are asked to take one of the two stacks they just sorted and re-sort that stack into halves. One of the two stacks from the second sorting--a quarter of the original items--is then arranged in some order such as most important/least important. If the entire unsorted stack has fewer than about twenty cards you might skip the second sort and have the respondents rank half of the

cards. For instance, if you have a stack of eighteen cards, ask people to sort them into two stacks of nine each and then rank one of the half-stacks of nine cards, without having them divide the stack of nine cards into two stacks of four or five cards before ranking.

This procedure is a shortened version of the process involved in asking someone to take a list of items and compare them pairwise for every possible pair. Data from pairwise comparisons will tell you a rank ordering of the items: the item which came out on top most often in the pairwise comparisons is ranked first, and so forth. A Q-sort accomplishes the same end with a slight loss of reliability but a great saving of time and effort. Of course, the Q-sort yields rankings only for the upper quarter of each respondent's list.

5.3 PRETESTING

After you have drafted your questionnaire, whenever possible pretest the instrument before you administer it to the group or the population for which it is intended. At a minimum, you need to have several--at least three--other people read the questionnaire to ascertain that all of the items and instructions are clearly worded. One or more of these reviewers should have the same characteristics as the group to which you will administer the instrument: similar age, educational level, and occupation.

If your instrument is designed to do little more than gather opinions from a small group, such as the participants in a workshop, then very little additional pretesting is necessary. However, if you plan to use the questionnaire repeatedly, or particularly if you plan to use the questionnaire to conduct a large survey, then you need to carefully consider the following concepts.

Validity and reliability. You can find a more complete discussion of validity and reliability in **section six.** Briefly, the **validity** of a questionnaire is the instrument's ability to measure what it purports to measure. Does each item in the instrument, and the instrument as a whole, measure the constructs and collect the information which you sought? A questionnaire item is a means of communicating between the person who wrote it and the person who responds to it. If the item fails to convey the question from the mind of the writer to the mind of the respondent, and to stimulate the respondent to answer with the desired information, then the item is not valid. Invalidity may arise because respondents misinterpret

the meaning of the item or because the person who interprets responses to an item assumes that a specific kind of response indicates something which it does not really mean. To avoid this problem, examine a copy of a draft of the questionnaire and imagine that you have compiled the responses from a group of people. Then, imagining varied response patterns, think about how you would interpret the meaning of each response pattern. Question your own interpretation as if you were a skeptical recipient of your ultimate evaluation report.

Reliability refers to whether or not you would get the same results if you administered an instrument to the same individual or group on two occasions with a reasonable amount of time between. How much a "reasonable amount of time" is depends entirely upon the kind of instrument. An intelligence test should yield the same IQ if administered to the same individual on two occasions several weeks apart. A questionnaire, on the other hand, may gather different results if administered a few hours apart because the individuals who respond may change their opinions. However, a questionnaire should gather the same results when administered with several days or weeks between administrations if no events or process which might change respondents' opinions occurred between testings.

An unreliable instrument cannot be valid; it cannot measure what it purports to measure if it is incapable of consistently measuring anything.

If you are undertaking a survey then you definitely need to test your questionnaire for reliability. To do that, you need to administer the questionniare to a subgroup from the population with which you intend to use the questionnaire and then correlate the results from the two administrations to the same group. It is always wise to correlate each item's results even if your instrument yields some form of total score reflecting a sum or average of the item scores. Correlations and correlation coefficients are discussed in **section nine.**

The generally accepted range for test-retest coefficients of reliability for questionnaires is around .80. **BUT,** if the test-retest reliability coefficient drops below .75, the reliability of the questionnaire instrument is seriously in doubt. You may obtain a test-retest reliability coefficient which is statistically significant for the number of data points which you have, but that significance per se is irrelevant to the reliability of the instrument.

Dummy tables. If you have questionnaire items which ask for demographic information such as age, sex, race, educational level, and the like, then you may thank yourself later if you prepare blank tables, charts, or graphs in advance and compare the data which the questionnaire asks the respondents to give with what you will need to complete your tables, etc. For example, you might ask respondents to reveal their age by simply asking, "How old are you?" and leaving them a blank space in which to write in a number. Or, you might ask for ages in the following manner:

How old were you on your last birthday?
_____18 or younger
_____19 to 21
_____22 to 25
_____26 to 30
_____30 to 35
_____36 or older

The data which you collect through the first question will allow you to calculate averages and other kinds of descriptive statistics, including a breakdown of the data in the same categories, which you would not get from the second sample. However, the data which you gather by asking respondents to give their age in categories does not enable you to do the reverse, to calculate averages and so forth.

A final note. While questionnaires are a frequently used and often invaluable tool for conducting needs assessment and for evaluating the content and process of an educational program, before you can use a questionnaire to evaluate impact or outcome you need to ascertain that the questionnaire validly and reliably measures outcome or impact. Tests are more appropriate measures of cognitive or psychomotor outcomes and impacts. You can use questionnaires to measure such things as changes in trainees' performance or skill level by administering instruments to people who have the capability of observing performance and skill. However, the data which such questionnaires collect may be subject to various and confounding interpretations unless you take care to ensure that the items truly measure the trainees' performance or skill. I would recommend that you seek the help of a qualified consultant for such endeavors.

5.4 INTERPRETING QUESTIONNAIRE RESULTS

Although **section nine: Analyzing Data** covers the basics of manipulating various data, there are several concepts related to interpreting and analyzing data from questionnaires which warrant attention here.

The common method of summarizing responses from Likert-type items involves assigning a numerical value to the response categories (e.g. strongly agree = 5; agree = 4;... strongly disagree = 1). Then analysts calculate measures of central tendency (see 9.2.3) and even estimate measures of dispersion (9.2.4). However, the response categories in a Likert-type item are categorical, not metrical; the data are ordinal rather than interval (see 9.1). Social scientists have debated for some time the merits and pitfalls of analyzing questionnaire data, particularly Likert-type data, as if it were interval data. Those in favor argue that the data almost achieve interval-level measurement and say that a great deal of information is lost when we strictly adhere to the laws of measurement and treat the data as ordinal. Opponents declare that an assumption of interval measurement compromises the interpretation of the information. I recommend case-by-case analysis of the meaning. In those instances when the median response differs greatly from the mean calculated by assigning numbers to the response categories, I obey the laws of measurement and attach the most significance to the median. When this happens, the distribution of responses is skewed (see 9.2.3). When no skewing appears--when the mean and median are identical or very nearly so--there is no problem worthy of genuine concern.

With Likert-type, rating scale, and semantic-differential items **which ask for opinions about people,** the expected response-pattern for adequate-to-good performances or sessions nearly always runs toward the higher end of the response spectrum (or toward the positive choices in semantic-differential pairs). Suppose your are analyzing the results of the following item:

> How would you rate the skills of the workshop leaders in terms of leading a group discussion? Please, circle one number on the scale below:
>
> 1 2 3 4 5 6 7 8 9
>
> Poor Fair Average Good Excellent

If the median or modal response falls below 7, I would interpret that to mean the respondents had a fairly low opinion of the leaders' discussion/leadership skills. A "7" would indicate relatively average skills, an "8" or "9" good-to-excellent. In short, you can expect the distribution of

responses to most questionnaire items asking for subjective opinions on people or their performance to come from a distribution with an expected average rating somewhere nearer to the positive end of the response spectrum rather than the literal middle. In the case of a semantic-differential item you should expect at least three fourths of the respondents to choose the positive descriptor in each pair; a good rating is one closer to ninety percent positive responses.

Exercise restraint against the temptation to infer cause-and-effect statements from questionnaire data. You might err if you decide that a particular response pattern on an questionnaire item implies a causative antecedent for the opinions represented by the pattern. Many thought processes, motivational considerations, and emotional shades combine to stimulate the behavior of marking a response to a questionnaire item. Hunches upon which to base future content or process choices are neither dangerous nor exceptional uses of questionnaire data; rather, they are the most common pupose of administering questionnaires to the participants in an educational program. The danger lies in attempting to draw scientific conclusions from the data and disseminating it to others who do not have ability to critique the data and the conclusions drawn from them.

The admonition I state at the end of 6.2 applies equally well here; **under no circumstances should this section of the manual be construed as either the necessary or sufficient instruction to enable the construction of an attitude measure capable of making diagnoses or of verifying changes after experimental research treatments.** Such endeavors require the efforts of trained psychometrists and scores, if not hundreds, of person-hours of work.

SELF-TEST ON SECTION FIVE

You will probably gain the most benefit from this self-test if you record your answers to these items on a separate sheet of paper and then compare your answers to the answers given in the pages following this self-test.

1. If two questionnaires are otherwise of equal quality but one is shorter than the other, the shorter instrument will have greater____.

2. Critique the following questionnaire items:

   a. Do you agree or disagree that the facilitators for this workshop encouraged questions and were knowledgeable in the subject matter covered?

   b. If you were going to select volunteers for doing methods raps, what were some of the circumspectly prescribed desiderata you would want to include?

   c. How would you change this workshop if you were going to attend it again with a fresh start? Please, write your answer in this space.

   d. Who promulgates guidelines and oversees the expenditure of funds under Titles X, XIX, and XX of the Public Health Services Act?

3. Define response sets and describe the "indecisive funnel."

4. Give three reasons why one is wise to begin constructing a questionnaire as soon as time permits.

5. Identify each of the following item types:

   a. Family planning educators are nonjudgmental about almost every personal quality in their peers and clients except about judgmentalness.

   strongly agree agree disagree strongly disagree

   b. How soon after their training ends should new education volunteers be sent into the field? Assume that they have two hours of practice-speaking weekly. They are ready in:

0	1	2	3	4	5	6	7	8	9
immediately		weeks		a month		weeks		2 months	longer than 2 months

   c. In each of the pairs of descriptors below, circle the one you would be most likely to use when describing your experiences here:

   exciting boring

   d. Among the various methods of contraception available, high blood pressure is a relative contraindication for_____.

e. I think family life education is...

6. What are the relative merits of offering a neutral response choice on a Likert-type item versus not offering one?

7. What happens when a questionnaire-writer abbreviates the response categories on a Likert-type with "1 2 3 4 5," or with the first five letters of the alphabet?

8. What is Q-sort exercise?

9. An acceptable reliablility coefficient for a questionnaire is_____ or above, but if the coefficient drops below_____some revisions or a complete rewrite is in order.

10. If you give your respondents a rating-scale item on a questionnaire which asks that they rate an educator's performance as a trainer on a scale from one (bad) to ten (excellent), about where would you expect the average response to fall on the ten-point scale for an average performance by the educator? Why?

11. What is the minimal amount of pretesting for a questionnaire in order to ensure some level of validity and reliability?

ANSWERS TO SELF-TEST ON SECTION FIVE

1. Reliability. Long questionnaires tend to cause fatigue and lead to less effort by respondents in considering and marking replies.

2. a. This item is two questions and should be divided two items. As it is written, you would not know if respondents were rating the knowledgeability of the facilitators or their encouragement of questions.

   b. The second "were" is grammatically incorrect (should be "are") and would lead respondents to "grade" the item. The vocabulary is stilted and awkward to the point of stuffiness.

   c. What does "fresh start" mean; am I, the respondent, to assume I am going to attend the same workshop over again as a new participant or as one who has been to it already? The space left below the item is much too small, leaving the impression that the writer does not really want the respondents' opinions.

   d. This question presumes a great deal of knowledge on the part of the respondent and will probably injure some egos and gather only a few responses.

3. Response sets occur when a respondent systematically marks the same response to a list of items, e.g. when one marks "strongly agree" to all of the items on a list of Likert-type items. An indecisive funnel occurs when a respondent begins answering a long list of Likert-type or similar items with strong opinions but tends toward less decisive answers further down the list, giving the appearances of a V- or funnel-shape to the response pattern.

4. To allow time for proofreading; checking aginst the perceptions of others, particularly laypersons from the population of interest; to permit university students time to pass it through their faculty and university committees; to allow time for pretesting for validity and reliability; to allow time for attractive preparation of the instrument.

5. a. Likert-type; b. rating scale; c. semantic differential; d. cloze; e. sentence completion.

6. Offering no neutral response choice helps avoid the funnel effect and does not allow those who genuinely have no opinion to mark a response, thus increasing the number of omitted items. Offering a neutral response choice creates a haven for the indecisive but reduces the number of omissions.

7. The reliability and validity drop. Research has shown that these practices cause respondents to attach connotations to the abbreviations which are irrelevant to the construct which they abbreviate.

8. A respondent receives a list of statements, words, or phrases, such as a list of possible topics for a workshop, printed on individual slips of paper or index cards with instructions to sort them, usually into

upper and lower halves according to some continuum (important/unimportant, good/bad...). Next, the respondent sorts the upper or lower half into halves, resulting in two stacks of one quarter of the original items. Then, the quarter of the original items of most interest to the persons who initiated the Q-sort is ranked by the respondent.

9. .80, but trouble is around the bend at .75.

10. About seven or eight. People will shy away from rating even the poorest trainer they have ever seen in the range of one, two, or three, thus moving the middle of the response continuum toward the more positive end of the rating scale.

11. Minimally (as in bare-bones minimum) have the items reviewed by at least three other people for content and clarity, at least one of whom should be representative of the group to whom the questionnaire will go in terms of educational level, reading ability, age, sex, and the like. Section 5.3 covers pretesting.

section six:
WRITING TESTS

After reading this section and learning the material in it you will be able to:

1. Identify sources of information on the currently available published tests.

2. Define reliability and describe methods to estimate reliability coefficients.

3. State acceptable levels of instrumental reliability.

4. Define and state how to estimate content, construct, concurrent, and predictive validity.

5. Construct a test by using a table of specifications.

6. Write test items of the following types:

   a. multiple-choice
   b. interpretive task
   c. matching
   d. cloze
   e. true-false
   f. dichotomous-choice
   g. hierarchical recognition
   h. essay

7. Score a test and interpret its results.

8. State four criteria for selecting published tests for use in the evaluation of educational programs.

If you believe that you already know this material and can perform the tasks outlined by the above objectives, then you may want to skip this section and go on to **section seven.** If you are not sure, or it has been a while, you may want to turn to the self-test in the back of this section to review or check what you have learned.

section six: WRITING TESTS

"That test wasn't multiple-choice--it was multiple-guess! You'd have to be able to read that #@¢!'s mind to figure out the correct answers!"

--One Trillion Students

Unfortunately, everyone has probably had the frustrating experience of taking a test which seemed unfair, unjust. Those kinds of tests are sad, not only because the students feel cheated and angry with the instructor, but also because the teacher doesn't really examine the test takers' knowledge--only their ability to read minds or make guesses.

The information and skills in this section will help you avoid such unfortunate abuses. This section begins with some important concepts about and limitations on the use of tests, then turns to specific steps for constructing tests.

6.1 EVALUATION VERSUS MEASUREMENT

When one sits down to construct a test, some framework is necessary to guide the process. Obviously, the uses of the test help determine the content of the test's questions (**items**). An intelligence test includes items based upon more kinds of skills and knowledge than a test covering plane geometry. But there could be a much more subtle, and significant, difference between the two tests. When a psychologist marks an IQ test, the raw score (usually the number of correct items minus a percentage of the wrong answers to correct for guessing) is compared to a table of **norms**. The number in the table corresponding to any raw score is the normed score. The normed score translates into an IQ. By itself, the raw score has no meaning until compared to the norm.

Thus, psychologists call IQ tests **normative referenced**, or **norm referenced** tests. They derive the norms by giving the items to thousands or tens of thousands of individuals and then applying statistical procedures to the results. Norms against which to define the meaning of a test's raw score may be established in several other ways.

The plane geometry test might also be norm referenced if the knowledge measured by the items comprises a representative sample of the body of facts called plane geometry. However, if the geometry test measured knowl-

edge against some other standard, such as the content of the lessons in a geometry course or the knowledge about geometry necessary to take a course in trigonometry, then psychologists call it **criterion referenced.**

If a geometry teacher took a list of items and constructed various tests consisting of subgroups of those items, then performed various kinds of analyses upon the results of administering them several times, the teacher could establish norms for the test. If, on the other hand, the teacher stated objectives at the beginning of the course and then used those objectives to construct test items, the resulting instrument would be criterion referenced.

There are significant differences between criterion referenced and normative referenced tests in terms of the meaning of their scores and their interpretation for evaluative purposes. Norm referenced tests are measurement tools; psychologists and educators use them as tools for diagnosing or predicting human knowledge, attitudes, performance, and behavior. If candidates earn a perfect score on a normative referenced test, then the instrument did not measure the candidates' attributes because they obviously know more than the test measured. Similarly, if candidates answer no items or only a few items correctly on a normative referenced test, then that test is inappropriate for those individuals. The former is called a **ceiling effect,** and the latter a **basement effect.**

Psychometrists, the professionals who study the science of measurement and test construction, go back to the drawing board with a norm referenced test which consistently suffers from ceiling or basement effects. However, someone constructing a criterion referenced test would probably be less concerned with the ceiling effect. If several individuals in a group score 100 percent on a criterion referenced test, then the people who constructed the test know that those with perfect scores have met the criteria, if their instrument is valid and reliable (see 6.2).

A criterion referenced instrument may be constructed with relatively little time, effort, or resources, compared to what it takes to construct a sound normative referenced test. Universities and corporations, such as the College Entrance Examination Board (CEEB), The American College Testing Service (ACTS), and the Psychological Corporation, devote several person-years and thousands of dollars to the process of constructing the instruments which they sell and administer. In addition, constructing achieve-

ment tests, such as a norm referenced plane geometry exam, is relatively easy compared to constructing personality inventories, vocational interest inventories, and attitudinal diagnostic tools. The **Minnesota Multiphasic Personality Inventory (MMPI)** took years to develop. Even though the MMPI is beyond doubt the most extensively examined and frequently administered tool for diagnosing personalities and their disorders, controversy about its reliability and validity persists.

This section is intended as a guide for use in constructing criterion referenced tests of cognitive ability and knowledge. If you follow the guidance offered in this section, the tests which you construct to evaluate the impact of educational programs should serve your purposes quite well. However, if you want to devise a **norm referenced** test of knowledge in some area of family plannning, you have a great deal more work to do after you write your items as advised here. If you do want to construct a norm referenced test designed to measure knowledge in some area of family planning, you can find the additional information you need in Pasquariella and Wishik's (1975) manual.

Under no circumstances should this section be construed as a guide for constructing diagnostic instruments, measures of attitude, or personality inventories. Section 6.6 describes criteria for selecting instruments which other people have developed and made available, as well as sources of information about other people's tests.

If you are interested in learning about psychometrics, test construction, or any other topic covered here, then you might refer to Thorndike and Hagen (1977), or Cronbach (1960). There are numerous other textbooks available, written at various levels of difficulty.

6.2 VALIDITY AND RELIABILITY

Simply stated, an invalid instrument is basically useless, and an unreliable instrument cannot be valid. If an instrument is either unreliable or invalid, then the data collected from administering it are subject to various interpretations: do the results permit useful evaluative conclusions about the educational program or do they simply confirm the poor quality of the test?

6.2.1 VALIDITY

An instrument has validity if it measures what it purports to measure. Psychometrists recognize four kinds of validity, each having differ-

ent applications and different significance depending upon the purposes of a given test: content, construct, concurrent, or predictive.

Content validity. Of the four, content validity will be of most frequent concern when constructing tests for evaluative purposes. A test has content validity if the items measure knowledge of a representative sample of all the knowledge in a given body of material. A test which emphasizes only portions of the total area it is supposed to cover, to the exclusion of other major portions, lacks content validity. Some teachers are fond of giving tests and examinations composed of items which heavily emphasize a small, sometimes obscure, part of the material or training program. You have taken such an invalid test if you walk way from the exam thinking, "Why did they ask so much about this and nothing about that, that, and that?"

Psychometrists establish content validity for normative referenced tests by convening committees of experts in the area and asking them to select a given number of items from a large pool of potential items. The panel usually receives two or three times as many items as will appear on the final instrument along with a **table of specifications** (see 6.3.1). The presumption is that the experts, based upon their knowledge of the field, will select a representative sample of items. Sometimes one test is examined for content validity by attempting to establish its concurrent validity with another test which has already been proven to have content validity.

Concurrent validity. Suppose you have a test, call it test A, which you know for a fact has validity. Test B may or may not have validity. If you administer the two tests to a group of appropriate subjects and then correlate their scores on A and B, a high correlation (see 9.3.2) indicates that test B has concurrent validity with test A. Since there are relatively few published tests covering the entire field of sexuality, to say nothing of knowledge in the area of family planning, I can envision only rare instances when concurrent validity would be of concern to a family planning educator.

Construct validity. A construct is some human attribute such as intelligence, personality, or attitude, or a concept such as coordination or mechanical aptitude. Construct validity centers around the fact that a person taking a test performs a specific behavior, i.e. marks responses to

the test's items. However, do those individual item-answering behaviors, or their sum total reflected in a test score, really measure intelligence, attitude, etc.? Although the first intelligence test was written and administered about eighty years ago, psychometrists and psychologists still argue the meaning of the construct "intelligence." An IQ score is in reality a statement about how a person compares with the mass of individuals of the same age. But the questions on the IQ test itself are subject to ongoing controversy because some claim that the very choice of words for an item is embedded in the culture from which the item content arose.

Psychometrists frequently attempt to establish construct validity through concurrent or predictive validity. In this instance, concurrent validity may be established against some criterion other than another test. For example, suppose you want to construct a test which will help you screen job applicants and determine which applicants will be good educators. The validity of such a test lies in its ability to foresee behaviors on the job. Thus, you might argue that your test has construct validity if its results correlate with behavioral observations (see **section eight**) of people who take it and then perform educational tasks.

Sometimes test constructors develop some theory about the construct they want to measure and then compare the test results with what they observed about people in relation to their theory. The **MMPI** was examined for construct validity by comparing its scores with diagnostic reports written by psychologists and psychiatrists who interviewed people who had taken the **MMPI.**

Predictive validity. Some texts refer to this concept as either **criterion-related, statistical,** or **analytical validity.** A test which is capable of accurately foreseeing future behavior or performance has predictive validity.

The best examples of tests developed for predictive validity purposes are the **Scholastic Aptitude Test (SAT), the American College Test (ACT)** and the various graduate and professional school exams, such as the **Graduate Record Examination (GRE)** and the **Medical College Admission Test (MCAT).** Of course, none of these tests will predict with perfect accuracy how well a student will do in college, or graduate or professional school. However, they give admissions officers normative data for comparing applicants with their peers. These tests have much higher predictive validity than meas-

ures such as grade point averages because grades reflect not only a student's achievement and aptitude, but also the standards of the teachers and institutions which assigned the grades.

You might be concerned with predictive validity in evaluation if you want to construct a test which will predict how trainees or students will perform in the future. If you conduct a workshop or course with the objective of imparting or improving skills, then you might test the skills which your trainees have learned. You might establish the content validity of your test by comparing it to the objectives which you set for the educational program. However, it has predictive validity only if you can demonstrate that the test scores in some way correlate with the trainees' future behavior.

You might do that by sending observers to watch them on the job to determine if they are executing certain tasks using the skills which you hoped to teach them, and then correlating your observational results with the scores on your test.

"Face validity." Most books on testing scoff at the concept of face validity, calling it meaningless. An instrument has face validity if it is valid "on its face"; if it is obvious to the reader that the instrument measures what it purports to measure.

Your questionnaires have face validity for those items which ask for straightforward information, information which is either accurately revealed or not. Questionnaire items which seek opinions, such as a workshop participant's perceptions of the workshop's content and process, have face validity if they are reliable and if their meaning is explicitly clear. Difficulties can arise if you assume that the data from a questionnaire imply more than what the question asks for on its face. For instance, questionnaire data which confirm that you have conducted a good educational program from the standpoint of content and process **do not** imply that you have had the desired outcome, impact, or both.

6.2.2 RELIABILITY

The reliability of an instrument refers to its consistency, the repeatability, of its results. If you administer the same instrument to an individual or group on two separate occasions and get the same results each time, then the instrument is reliable.

Of course, factors irrelevant to the test itself will cause some variation in the scores. Respondents make mechanical errors when they record their answers (e.g., calculate the correct answer for a mathematical test item and then record it incorrectly). They get distracted by noise, temperature, and stress. However, an instrument can add to these random errors. Testing an instrument's reliability involves an attempt to ascertain that the combination of random errors and errors inherent in the test is kept to a minimum.

But what is a minimum? Reliability is expressed in terms of correlation coefficients (see 9.3.2). These coefficients vary from -1.00 to +1.00. Coefficients indicate higher reliability the closer they get to +1.00.

Methods of determining reliability. Psychometrists recognize three common methods for determining the reliability of a test: **test-retest, split-halves, and parallel forms.**

If you want to test the reliability of an instrument, you will need to learn the information in **section nine** regarding correlation coefficents, or to enlist the assistance of someone who can calculate and interpret correlations.

If you administer a test to the same group of individuals on two occasions a few days or weeks apart and correlate their scores, the resulting correlation coefficient is the **test-retest** reliability estimate. Of course, some variation in the test scores from one administration to another will result because the subjects who take the test become familiar with the items during the first administration. To reduce the chances of this phenomenon, you need to separate the two administrations by as much time as possible and avoid revealing to the subjects that they will be taking the same test again. Obviously, you cannot perform a test-retest reliability study on an instrument if you give it as a pretest and a posttest with some instruction sandwiched in between.

To use the **split-halves method,** construct your test and administer it to as large a group as possible (no less than twenty) from the population to whom you plan to administer the test. Then score the items and divide the test into two equal halves. Use the odd and even items as the halves whenever possible; using the first and second halves is a bad practice. Correlate the scores for the two halves for each person who took the test

to yield a split-halves correlation (see 9.3.2) That correlation yields a conservative (low) estimate of the test's reliability. (See the Spearman-Brown Prophesy Formula, discussed below.)

The split-halves method of checking reliability takes less time and effort but is less accurate. However, if you obtain an acceptable reliability coefficient from the split-halves method, then you can be sure that you would also obtain an acceptable coefficient with the test-retest method, because the split-halves methods yields a lower correlation coefficient. Therefore, you can safely check a new test's reliability by planning to conduct a test-retest reliability study, administering it once, and calculating the split-halves reliability estimate; you may omit the second administration of the test (the "retest") if the split-halves reliability coefficient is acceptable.

However, the two halves of the instrument must truly be halves in the sense of equal parts. Problems with this requirement commonly occur for one of two reasons. Psychometrists call an instrument which is timed a **speed test** because an individual's score depends not only upon knowledge but also upon ability to work quickly or under stress. You administer a **power test** when you give candidates as much time as they need, within reason, to complete the test. Split-halves reliability estimates cannot always be interpreted as accurate reliability indicators for speed tests. You do not need to be concerned with this problem, even if you set time limits for taking the test, as long as the test fits the allotted time. That is, as long as the length of the test permits completion of it by a vast majority of candidates in the allotted time.

The second problem which can arise with the halves occurs when the likelihood that a candidate will correctly answer certain items varies from item to item. Obviously, one has a greater chance of guessing a correct answer to a true-false item than to a multiple-choice item, which is, in turn, easier than a cloze item (see 6.3.2). Therefore, if you have different kinds of items on a test, the split-halves method of estimating reliability will work if you have the same number of each kind of item in each half. In addition, if all of the items on a test are of the same type, but some are easier than others, you can encounter problems with the split-halves reliability estimate. If all of your items are of the same type, but your candidates are consistently scoring much higher on one half than

on the other, a split-halves reliability estimate would probably be in error.

Measurement experts have proven that increasing the number of items on a test within reasonable limits will improve the reliability of the instrument. Since a split-halves reliability coefficient is actually based upon half of the items, the reliability of the whole test will be higher. There is a formula to apply to split-halves reliability coefficients to estimate the reliability of the whole test, the **Spearman-Brown Prophecy Formula:**

$$r_{11} = \frac{2(r_{\frac{1}{2}\frac{1}{2}})}{1 + r_{\frac{1}{2}\frac{1}{2}}}$$

In the formula, r_{11} is the estimate for the reliability coefficent of the entire instrument and $r_{\frac{1}{2}\frac{1}{2}}$ (r half) is the reliability coefficent which you obtained from the split-halves reliability estimating procedure. Thus, if the correlation of the two halves' scores for a group is .70, you would get a Spearman-Brown Prophecy Formula estimate of:

$$r_{11} = \frac{2(.70)}{1 + .70} = \frac{1.40}{1.70} = .82$$

The result of .82 represents an approximation of the actual reliability of the total instrument; all one can say for certain is that the true reliability lies someplace between .70 and .82. There is good evidence that the true reliability is closer to the higher number.

The **parallel forms** method of estimating instruments' reliability requires that you contruct two forms of the same test, presumably containing the same number of items covering the same body of material and of equal difficulty. Then you administer the two parallel forms to one group of individuals, either one after the other on the same occasion or on two separate occasions. The correlation of the individuals' scores on the two forms gives you an estimate of the reliability for each form. If you decide to administer both forms to one group at one sitting, mix the order in which you administer the forms; give half of the group one form first and the other half the second form, then reverse the order.

If you find an instrument has a reliability which is too low, you can increase it in several ways. In addition to lengthening the instrument, you can examine the items and try to identify those which contributed most to the poor reliability. When you examine the items, you are looking for items which large numbers of individuals answered incorrectly, an indication that there was a lot of guessing. (See 6.5.2.)

Items which a large number of people answered correctly increase reliability falsely; they inflate the true reliability of the test because an easy item makes it more likely that people will answer correctly twice. Similarly, if you have a ceiling effect, your reliability coefficient for the entire test will be inflated. A basement effect will create a similarly false, and inflated, estimate of reliability. Consequently, if you are constructing a criterion referenced test based upon specific objectives, and are therefore not concerned about ceiling effect when you use the instrument as a posttest, then you should perform your reliability study using either the split-halves or the parallel forms method and give the instrument as a pretest.

Poorly worded, ambiguous, trivial, and extremely difficult items reduce the reliability of a test and should undergo revision.

Acceptable levels of reliability. There is no way to establish an objective standard against which to compare the reliability of instruments. Like the question "How wide is narrow?" the answer is relative. For questionnaires, acceptable reliability coefficients for the entire instrument are .80 and higher, but not less. (A coefficient of .70 is not almost as good as .80 -- .70 implies a 25 percent drop in reliabilty from .80!) You would be wise when checking the reliability of a questionnaire to use the test-retest method. The split-halves method is inappropriate, and the parallel forms method is too cumbersome. Then, with your test-retest data, estimate the reliability of **each item** and revise those items with low test-retest correlation coefficients.

Test-retest reliability coefficients for individual questionnaire items may fall into the range of .65 or even .60. If they fall lower, they should be revised. To estimate the reliability of an entire questionnaire, follow this procedure. Consult a statistics text which has a table for translating correlation coefficients into a statistic called **Fischer's z.** Then calculate the average of those Fishcher's z's and use the table to

convert that average z back into a correlation coefficient. The result is an estimate of the reliability of the entire questionnaire. Do **not** average the test-retest correlations of the individual items; the result would not be accurate or proper.

For tests of knowledge, reliability coefficients should be at least .85; professionals will not accept reliabilities below .90 or even .95 for standardized tests. For locally constructed tests of knowledge, aim at a reliability of about .90, be satisfied with .85, and do some revising if it falls below .85.

A final comment. Why is reliability so important? Suppose you construct or adopt a test and give it as a pretest and posttest. If the test has poor reliability then scores will vary from pre- to posttest because of that poor reliability--thus muddling your attempts to interpret the effects which your program had upon the participants. If you had expected or hoped for changes and there were none, is that because your program had no effect or because the test could not reliably measure the effects? An unreliable test is a lot like an elastic ruler which measures a yard as 29 inches today and 44 inches tomorrow--you never know in any instance how much error is involved or in which direction.

6.3 CONSTRUCTION OF A TEST

If you follow the steps described here you will greatly increase your chances of constructing valid and reliable tests of knowledge:

1. **Write a table of specifications.**
2. **Determine the number of items and the type(s) of items.**
3. **Create an item pool.**
4. **Select the items with attention to validity requirements.**
5. **Prepare the test papers/booklets and answer sheets.**
6. **Check the reliability and revise as needed based upon item analysis.**

A table of specifications lists the topic areas which the test will cover, along with their relative emphasis in the test. Then you determine the number of items you will include in the test based upon the relative difficulty of various item types, the amount of time you will allot for the test, and the amount of time each item is likely to require.

Next, you draft two or three times as many items as you are planning to include in the test, an **item pool.** From the item pool you select items for inclusion in the instrument, with attention to the importance of test validity.

Finally, you prepare the test, and preferably an answer sheet, as early as possible. Of course, you will need to estimate the reliability of the test using one of the methods outlined in 6.2. After your first one or two administrations of a new test, you may find that you need to make revisions, but a well-constructed test can be used over and over again.

All of the guidelines which I mentioned in subsection 5.1 apply to constructing tests as well as questionnaires, with two significant exceptions. Since the purpose of writing a test is to measure knowledge, you can, of course, make assumptions and presumptions about what test takers know.

The other guideline which does not apply here is the one suggesting that you mix item types in questionnaires. Since the reliability of a test is most easily estimated using the split-halves procedure described in 6.2, the need for the two halves of a test to be equal sometimes necessitates limiting a test to one type of item. Sometimes a mixture of items is necessary to measure knowledge of certain facts. Some mixing of item types is also desirable on a test, although not to the extent which I recommend in questionnaires, because changing item types keeps test takers more alert, therefore reducing random error and increasing reliability. (See split-halves in 6.2.2.)

6.3.1 TABLES OF SPECIFICATIONS

A table of specifications lists the topics which the test is designed to cover and the percentage of the test which will cover each topic. The percentages of items devoted to each topic area should parallel the percentages of time, effort, or homework in the session, course, or workshop in which the participants learned the material covered by the test.

If you write a list of explicit behavioral "outcome" objectives which specify the knowledge you want participants to have when they leave an educational program, then those objectives will not only serve as a guide for the participants during the educational program, but also suggest the topic areas which will appear in the table of specifications as well as specific items for the test.

The following is an example of a table of specifications for a test at the end of a four-hour seminar on the basics of contraception for new family planning clinic workers. The first column in the table lists the number of minutes devoted to each topic, excluding any breaks, introductory

remarks, or the like. The second column shows the percentage of the total 245 minutes (or four hours and five minutes) represented by the time in the first column. The last column, on the right, lists the percentage of items which the exam will include on each of the topics. (Section 9.2.2 describes percentage calculations.)

Table A. SAMPLE TABLE OF SPECIFICATIONS

TOPICS:	Amount of time spent	% of time spent	% of items
Anatomy & physiology of contraception	60 min	25%	25%
Contraceptive effectiveness rules	15	6	5
Prescription methods of contraception:			
Orals	45	19	20
IUDs	35	15	15
Diaphragms	25	10	10
Sub-total, prescription methods	105 min	44%	45%
Non-prescription methods:			
Spermicides	15	6	5
Condoms	15	6	5
Natural FP/Fertility Awareness	25	10	10
Sub-total, non-prescription	55 min	22%	20%
Non-methods (douching & withdrawal)	10	4	5
TOTALS	245 min	101%*	100%

*Exceeds 100% due to rounding.

Note that the percentage of items in the right-hand column closely approximates the percentages of time spent on each of the topics included in the session. These similarities do not have to be absolute. Sometimes you will find it necessary to place five to ten percent more emphasis in terms of the number of test items for a particular topic than the emphasis which that topic received during the educational program itself. You might do that because you feel that particular material to be of extreme importance, because more items are needed to adequately cover the material, or for other considerations which will become clear later.

You might also base the percentage of items on each topic in a table of specifications upon considerations other than the amount of time spent on each topic. For example, you might decide to include however many items

it takes to measure mastery of the knowledge and skills specified in your behavioral objectives for the learning process. When psychometrists construct norm referenced achievement tests, they devise tables of specifications based upon expert opinion about the percentages which various topics represent of the total body of knowledge which the achievement test measures. Thus, if you intend to construct a test to use as a screening device for job applicants or for applicants who want to volunteer as educators, you might write an outline of the material you want to measure with the test and, with the assistance of peers/experts, construct a table of specifications with the outline as a guide. A test constructed to measure new clinic-volunteer applicants' knowledge of the information which they need to perform various functions in your family planning clinic might be based upon a list of facts and skills derived from interviews with the clinic staff members who will supervise the volunteers.

6.3.2 DETERMINING THE NUMBER AND TYPES OF ITEMS

With an adequate table of specifications, you are ready to decide how many items you will include in the test and what types of items you will write. You will, of course, need to consider the reading level at which the test takers function, the kinds of items to which the material lends itself, and the considerations regarding split-halves reliability coefficient estimates discussed in 6.2.2.

However, the two major considerations which will guide your final decisions are (a) the amount of time you wish to devote to administering the test and (b) the relative difficulty of the items. Thorndike and Hagan (1977, p. 212) state that "as a rough rule of thumb, the typical student might require from 30 to 45 seconds to read and attempt to answer a simple factual type multiple-choice or true-false item and from 75 to 100 seconds to read and attempt a fairly complex multiple-choice item requiring problem solving." Based upon those estimates and my own experience with writing tests, I have been able to roughly estimate that matching items require approximately 30 seconds per match. Cloze items require widely varying amounts of time, but I usually estimate 30 seconds per cloze item. Interpretive task items, particularly those requiring that students examine and interpret graphs, charts, or the like, require approximately three times as much time to complete as it takes to read the interpretive material and the items about that material. Hierarchical recognition items follow the same

approximate pattern as matching items. Of course, the amount of time required to respond to an essay item varies immensely. However, you can get an approximation by specifying in the essay item the amount of time, or the number of words, which you expect in answer to the item.

Some types of items are more difficult than others in terms of the percentage of a group expected to correctly answer an item. A smaller percentage of test takers will correctly answer a multiple-choice item than a true-false item, if for no other reason than that there is a greater chance of correctly guessing the response to a true-false item. If we assume that all other things are equal (reading level, content difficulty, student preparation...), then Lord (1952) estimated the following **average level of difficulty** for various item types in terms of the percentage of test takers we could expect to correctly answer each item type:

Item type:	Relative difficulty (%)
True-false, dichotomous-choice	85
Multiple-choice:	
3 response choices	77
4 response choices	74
5 response choices	70
Cloze, sentence completion	50

If we can expect 85 percent correct answers to the typical true-false item, then the validity of a true-false item's measurement of a given fact is seriously in doubt; what portion of that 85 percent represents trainees who really knew the material and what portion represents good or just lucky guesses? You can even out the effects of different item difficulties to a certain extent by applying corrections for guessing (see 6.5.3.) However, some consideration for the difficulty of the items must be included in your choice of item types in order to ensure that the test truly measures as much knowledge but as little test-taking and guessing ability as possible.

Assigning different point values to correct answers for different types of items will also help to reduce the influence of relative item difficulty. Based upon Lord's relative difficulties, if a true-false item is worth 1 point, then a multiple-choice item should be worth 2 to 3 points and a cloze or sentence completion item 4 points. If you choose this method, then the percentages in your table of specifications should be

interpreted to indicate the percentage of points on the overall test, not the number of items. Suppose, for instance, that you value true-false items at 1 point and multiple choice at 3 points. Now suppose that 15 percent of your test covers a given topic, and that the entire test is worth 60 points. Then the items covering that topic should be worth 15 percent of 60 points, or 9 points. So, you might write 3 multiple choice questions worth 3 points each to equal the required 9 points, or 9 true-false items worth 1 point each, or some combination of multiple-choice and true-false items totaling 9 points.

As a general rule, I shy away from writing tests which include only true-false items. I recommend tests composed mostly of multiple-choice, matching, and cloze items. These item types lend themselves well to most kinds of factual material. When other item types will more economically and accurately reflect knowledge of certain facts, I do not hesitate to include them.

In any case, if you plan to attempt a split-halves reliability estimate, each half of the test, whether you divide it by odd and even numbered items, by subsets of items, or by some other method, should have approximately the same number of items at each level of difficulty.

6.3.3 WRITING AND CHOOSING ITEMS

Having written your table of specifications and determined the number and the types of items you plan to include in your test, you will then know approximately how many items of each type you will need to write on each topic in the table of specifications. (Procedures for calculating percentages may be found in 9.2.2.)

Return to the example given in **Table A.** If I were to decide that I want to devote 45 minutes to testing the trainees' knowledge of contraception using exclusively multiple-choice items, then I would have time to administer approximately 60 items. Each multiple-choice item requires about 45 seconds to complete. Thus, I have 45 minutes at 60 seconds each, or 2,700 seconds divided by 45 seconds per item yields time for 60 items. Therefore, if the first topic in the table of specifications--anatomy and physiology of contraception--is to consume 25 percent of the test items, the instrument which I will administer should contain 15 multiple-choice items on the anatomy and physiology of contraception (25 percent of 60 is 15). Similarly, I would arrive at the number of items I need on the final instrument for each of the various topics.

Then create an **item pool**, a group of potential items from which to select the items for the final instrument. You should have sufficient items in the pool to give yourself a reasonably varied selection. As a rule of thumb, if you need five or fewer items on a given topic, attempt to write a pool of two and a half to three times the number of items you will need. In cases where you need six or more items on a given topic, try to write about twice as many as you will ultimately need. Rather than attempting to write a dozen perfect items, "brainstorm" items as quickly as possible without stopping to edit. Sometimes I write the stems of multiple-choice items and only one or two of what will ultimately be four or five response choices, leaving the rest for later.

Creating an item pool for a test serves several purposes. If you are concerned about content validity, and want to allow a group of experts to select your items, then they will need a pool from which to select items representing a good cross section of the domain of knowledge involved. Every pool contains some bad items, unsalvageable even by good editing. Without an item pool some of those bad items would appear on the final instrument.

Even if content validity is based upon objectives specifying the mastery level of knowledge for the test (you do not plan to have a group of experts select the items), you still need to have several people read the item pool to help you discover poorly worded or otherwise bad items. Avoid items to which more than one response is defensibly correct. **The correct response to each item should be verified by reference to some authoritative source** even in areas where you are certain of your knowledge. You can probably easily recall a situation in which someone wrote a test and one of the students won an argument with the instructor as to the correct response.

An item pool also serves as a source of questions for similar tests in the future, as well as an indispensable tool if you plan to check reliability using the parallel-forms method.

6.3.4 PREPARING THE TEST PAPERS

When you have completed the creation of your item pool and selected the items for inclusion on the test, you should prepare the test as far in advance as possible. You will also find it easier to grade the items if you give the test takers prepared answer-sheets even if there are only numbered spaces in which to record responses. (See 6.5.1.)

With the items you have selected from the pool, typically recorded on index cards, you face the task of deciding in which order to list the items on the test. Some test writers list the items in the same order as the chronology of the material covered. Others use reverse chronological order or random order. Studies have shown that correct chronological order will increase the total scores on a test by a small but statistically significant amount over random order or reverse chronological order. Listing the items in chronological order assists the trainees' recall process and helps to reinforce their learning. Reverse chronological or random order more finely distinguishes between those tests takers who truly know the material and those who know slightly less.

Of course, the final version of the test should be neatly typewritten with ample space for easy reading, and it should be carefully proofread for spelling, grammar, punctuation, numbering, and to ensure that no items from the final selection have been omitted from the test. By this point, those who have worked to write and select items and those who have typed the items will have difficulty finding errors because of their familiarity with the material. Consequently, someone who has not been extensively involved in the process to this point should be enlisted as an additional proofreader.

6.4 ITEM TYPES

The following catalogue of possible item types will probably contain few, if any, types of items which you have not seen many times on tests. However, you will probably find the information offered as guidance in constructing the various item types to be helpful when creating items and selecting items from a pool.

If you will both prepare a test and participate as an instructor in the sessions covering the material to be tested, take care that you do not use the same examples, case histories, or homework exercises during the instruction and again as items on the test. This becomes particularly important in situations when you wish to teach people how to apply knowledge or principles stated in general terms to specific situations.

6.4.1 MULTIPLE-CHOICE

A multiple-choice test item consists of a question, called a **stem**, with at least three and usually no more than five lettered responses to choose among.

Experienced psychometrists who write multiple-choice items have agreed upon several basic principles with respect to constructing both the stem and response choices. The first and overall guiding principle encourages the inclusion of **as much information in the stem as possible,** so as to keep the response choices as succinct and short as possible:

Sample 1. The testes major functions are:

A. producing both testosterone and spermatozoa.
B. producing both androgens and spermatozoa.
C. producing both spermatozoa and follicle stimulating hormone.
D. producing both spermatozoa and leukocytes (white blood cells).

Sample 1 is poor for two important reasons. Not only are the four response choices too long, but they are poorly selected. Each of the four response choices contains the word "spermatozoa," making all four responses the same in that respect. Sample 2 below corrects the first problem with sample item 1.

Sample 2. The two major functions of the testes are the production of both spermatozoa and

A. testosterone.
B. androgens.
C. follicle stimulating hormone.
D. leukocytes (white blood cells).

Sample item 2 still has some major difficulties. Since testosterone is one of the androgens, the male sex hormones, either answer A or answer B is technically correct for sample item 2. Alternative D is a poor response choice because it is highly unlikely to be the correct answer, a fact easily detectable by those who have studied human physiology. Sample item 3 further improves upon samples 1 and 2.

Sample 3. The two major functions of the testes are the production of both spermatozoa and

A. testosterone (androgens).
B. interstitial cell-stimulating hormone (ICSH).
C. follical stimulating hormone (FSH).
D. seminal mucus secreting hormone (SMSH).

All four of the response choices are now plausible even though response D is a nonexistent figment of my imagination.

The requirement that **good multiple-choice items have one indisputably correct response choice and that the alternatives offer plausible answers** proves indispensable to validity and reliability requirements and yet makes writing multiple-choice items difficult. Sometimes you simply have to omit an item from a test after you have administered it a few times because this basic principle proves impossible to fulfill. (See **Item Analysis,** 6.5.2.)

Avoid writing "trick" items and items which give hints or clear clues to the correct response in the stem. Some test writers have a fondness for including extraneous and irrelevant information in the stem, hoping that they will increase item difficulty by doing so. In fact, such practices reduce reliability and content validity. The stem of a multiple-choice item should pose a clear statement requiring that the respondents recall certain information, yet it should not reveal or give clues to the answer for other items on the same test.

Consider sample item 4.

> Sample 4. In the past, methods of contraception requiring periodic abstinence relied upon counting cycle length. Natural family planning or fertility awareness relies upon all of the following except:
>
> A. cervical mucus changes.
> B. thermal shifts.
> C. changes in sexual desire.
> D. breast tenderness.

This sample is poor because the first sentence in the stem is extraneous. Sample item 4 is also bad because "natural family planning" and "fertility awareness" are not correctly used as synonyms; many proponents of the Billings method call that method "natural family planning" although they base it solely upon cervical mucus changes.

In sample item 5 below, the word "womb" in the stem eliminates response choices B and D for any person taking the test who knows that womb is a synonym for "uterus":

> Sample 5. Objects made of plastic, metal, or both which are placed in the womb to prevent pregnancy are called:
>
> A. interuterine devices.
> B. interurinary devices.
> C. intrauterine devices.
> D. intraurinary devices.

Some multiple-choice items contain response choices such as "none of the above," "all of the above" ("sometimes the above," "on Thursdays none of the above," "maybe some of the above but never all of the above,"...). **Offering "all of the above" as a response choice typically gives away the answer** to those who have some but not necessarily all of the knowledge which the item seeks to measure:

> Sample 6. Which of the following are relative contraindications to prescribing oral contraceptives for an average patient under thirty years of age?
>
> A. history of liver disease
> B. history of high blood pressure
> C. history of smoking cigarettes
> D. all of the above

Any person who is aware that any pair of response choices A, B, or C is correct automatically knows that D, "all of the above," is the correct response. There is a grave temptation to use the response option "all of the above" in instances where you wish to determine if test takers are familiar with a list of things, such as the contraindications to birth control pills.

However, an item such as sample 4 which gives a list and **asks that the respondents find an exception** can solve this problem if the correct answer, the exception, is a plausible member of the list in question.

Using this technique of asking people to find an exception is also a good technique for avoiding negatives in multiple-choice item stems. Using negatives in any kind of test item bogs down students in semantics and logic, rather than having them concentrate on the material they need to recall correctly in order to respond to the item.

Offering a response choice of **"none of the above" suffers from the same drawbacks** as the use of "all of the above." Any respondent who can eliminate at least two of the other alternatives knows that "none of the above" must be correct.

Try to limit use of these phrases to those situtations in which you are at a total loss to find a plausible incorrect alternative. Even then, an item such as the following one offers a less reliable but still prefer-

able option, although requiring more than the usual amount of time to answer.

Sample 7. Which of the following will shorten the shelf-life of condoms?

1. storing condoms above 90 degrees F
2. storing them below 50 degrees F
3. storing them in a man's wallet
4. storing them in the glove department of a vehicle

A. 1,2, and 3
B. 2,3, and 4
C. 1,2, and 4
D. 1,3, and 4
E. 1,2,3, and 4

With an item like sample 7, people who know that one or two of the numbered phrases are correct still have difficulty with this item compared to those who know the correct answer, E.

However, the following revision of sample 7 asks for the same information, avoids the use of "none of the above," and is less complicated:

Sample 7, revised: Which of the following is an acceptable condition for storing condoms:

A. above 90° F
B. below 50° F
C. in a purse
D. in a man's wallet
E. in the glove compartment of a vehicle

6.4.2. TRUE-FALSE

Items which offer two response choices are called dichotomous-choice items. True-false items are one kind of such items. Some texts refer to this class of items as alternative-response. Other kinds of dichotomous-choice items appear in 6.4.3.

As with any item type, there are certain guidelines which you can follow to improve the validity and reliability of true-false items when you choose to use them. Foremost among these is the notion that a statement in the form of a **true-false item must obey the laws of bidentate logic; the statement must be absolutely correct or the opposite.** In order to ensure that a statement is either absolutely true or false, you need to be very familiar with the material the item covers, or to rely upon the knowledge of someone who is that familiar.

The first guideline almost always precludes **the appearance of such words as "all," "always," and "never,"** as well as terms or words which imply options: **"sometimes," "frequently," "may,"....**

Sample 8.

A. Birth control pills should never be given to women who have had tubal ligations.
B. Many women have headaches while taking birth control pills.
C. No woman should use the the same diaphragm for longer than two years.

In sample 8, item A is a false statement because birth control pills are sometimes prescribed for reasons other than contraception. However, a beginning family planning worker or even an experienced one may not be aware of that fact. In sample 8, item B, how many women is "many women"? Obviously, the answer to that item is a matter of opinion and neither true nor false. Item C in sample 8 is not only poorly worded but also a poor way to test knowledge of the material to which it alludes. In all three cases the items need to be reworded or changed to a different type.

Using negatives in true-false items, and particularly double negatives, has proven a poor practice. Putting a double negative in a true-false item requires more effort to decipher the meaning of a true or false response than should be required of a person taking the test and leads to confusion, loss of reliability, and unnecessary analysis of poor responses.

My admonition in section five to make sure that each item contains only one question also applies to true-false items. It is a bad practice to write some true-false items which are very short and some which are very long; **lists of true-false items should be approximately the same length whenever possible.** Otherwise, candidates attempting to respond to the items make suppositions about the correct response based upon the length of an item; research has shown that short items are presumed to be true and long items presumed to be false by some respondents while others reverse that thinking.

6.4.3 DICHOTOMOUS-CHOICE

When you give test takers a statement and ask them to mark one of two responses, you have given them a **dichotomous-choice,** or **alternative-choice,** item. A dichotomous-choice item is nothing more than a multiple-

choice item with only two response categories from which to choose. I have found dichotomous-choice items to be of limited value for the same reasons that I find true-false items to have limited value. Both item types have relative difficulties at the lowest end of the scale.

6.4.4 CLOZE ITEMS

These items are sometimes called completion or sentence completion items because they require the respondent to "fill in the blank." Writing good cloze items requires a great deal of thought because it is essential that one and only one correct word or phrase fit the blank.

Sample 9.

A. A patient who cannot take oral contraceptives and who has expelled two different IUDs should be offered _____ for contraception.
B. Storing _____ in the glove compartment of a car will seriously compromise their contraceptive effectiveness.

In sample 9, item A might be correctly answered by putting several different things into the blank space. However, in sample 9, item B, "condoms" is clearly the correct answer because no other form of contraception rationally completes that sentence.

All of the other guidelines regarding clarity, succinctness, grammar, diction, syntax, and so forth apply to cloze items as well as to other types.

6.4.5 MATCHING ITEMS

The sample below will illustrate the various principles involved in writing a good matching item:

Sample 10: A set of matching items:

Contraceptive methods:	Modes of action:
_____1. birth control pills	A. presents barrier to spermicide over cervix
_____2. diaphragm	B. unknown; probably prevents endometrial buildup
_____3. IUD	C. kills sperm in uterus
_____4. suppository, foam	D. kills sperm in vagina
	E. prevents ovulation

There is one more possible response than there are items to match to prevent respondents from correctly answering all four items by process of elimination if they know the answers to three. The response choice which

does not match to any of the items, choice C, is nonetheless a plausible choice. **Each of the two columns has a heading** which enables reference to the two columns in the instructions you give to the test takers. The instructions would read something like, "Choose the correct response to each of the four items listed in the column under contraceptive methods. Write the letter of the correct mode of action in the space provided." **Tell the respondents whether or not the same response can match to more than one item.** Otherwise, questions about that inevitably arise during the process of taking the test. Notice, however, that instructions do tell the respondents upon what basis they are to match the items in one column with responses in the other.

One could improve this matching item by switching the two columns and asking the respondents to match the methods to the mode of action. That will improve the item because the respondents will read the response column several times but probably only read the column containing the items once or twice. Reducing their amount of reading will shorten the amount of time it takes them to answer the items and improve reliability.

I shy away from the practice of having more than one response match an item. Research indicates that people become tied up in trying to decide how many of which match to what, rather than thinking about the content.

6.4.6 HIERARCHICAL RECOGNITION

What?

This kind of item requires that the respondent rank a list of things from most important to least important, most effective to least effective, or some similar construct. A **hierarchical recognition item is really a matching item in which the items are to be matched to numbers, letters, or some other form of rank.** If you asked test takers to rank the four contraceptive methods given in sample 10 above according to their use-effectiveness, you would have a hierarchical recognition item. This is, however, a perfect example of a place where a test item must be written and scored with reference to some authority because there are, of course, many different opinions about the use-effectiveness of the various methods listed in sample 10.

You might also present a list of things in narrative form--typed in lines across the page--and then ask the candidates to rewrite the list in the column of blanks which you provide. In either case, **try to avoid**

listing the items in proper order. Rely upon random or alphabetical order as a way to camouflage the correct ranking.

6.4.7 INTERPRETIVE TASK

Sometimes test writers give their respondents a few paragraphs, a chart, a graph, a map, a diagram, or some similar body of factual material and ask questions about the material. The items related to that material are called interpretive task items.

Some interpretive task items are of the matching type, such as those used in the common practice of requiring respondents to match the names of various reproductive organs with their location on a line drawing. Answering interpretive task items typically requires more thought and more analysis of the material than does simply locating places on a diagram, map, or chart.

Although interpretive task items might seem to consume a great deal of space at first glance, they can cover several different concepts and a lot of material in a relatively small space. For instance, you might present your respondents with a line drawing of the female reproductive anatomy and ask them to match what you labeled on the diagram with events or functions such as ovulation, conception, implantation, or the sites where various contraceptives act to prevent pregnancy. The items which follow a body of material assigned for an interpretive task may be of any type but are most commonly multiple-choice or matching type items.

6.5 SCORING AND INTERPRETING

This section discusses concepts related to scoring tests and interpreting the results for the purposes of revising and improving tests and for purposes of instructional improvement and development.

6.5.1 MECHANICS OF SCORING

Whenever possible you will find it to your benefit to **create an answer sheet for your test takers to use.** An answer sheet greatly increases the accuracy of scoring, reduces the amount of time it takes to score a test, and reduces the necessity of making more copies of the test when you administer it repeatedly. The answer sheet does not have to be fancy or formal; a simple, clearly typed sheet which enables test takers to circle or mark their responses should suffice. It may be easiest to give instructions like the following, as opposed to having people write their answers in a blank space or make check marks in blank spaces before the letters or numbers which represent the responses.

Mark your answers to this test on the answer sheet by putting an X over the answer which you think is the best response to each question. If you decide to change your answer, erase the first answer which you recorded as completely as possible. Please, do NOT make any marks on the test paper.

Example: If you think the correct answer is C, your sheet for that question should look like this:

A B X D E

With an **answer sheet arranged in the above format you can easily create a scoring key** by taking one of the answer sheets and using a razor blade or an art knife to cut holes through the sheet corresponding to the correct answers. Then you can mark an answer sheet by placing some distinctively colored mark on each candidate's answer sheet where the holes are not filled by an X when the key is laid over an answer sheet. Not only is this process quicker but it also reduces the possibility of scoring errors inherent in transferring eyesight back and forth from an answer sheet to a key list of correct answers.

Taking care in the scoring process will reduce random errors and increase reliability. Random errors in scoring are those which result from clerical errors in the scoring process, thus creating erroneous results for individual candidates.

Using such scoring keys makes it possible to administer a test and score the vast majority of the tests during the testing period. Some candidates inevitably finish earlier than others, and you can grade their tests while the other candidates complete their work. This gives the opportunity for almost instant feedback at the time when the test takers' curiosity about the correct answer is highest.

6.5.2 ITEM ANALYSIS

Using prepared answer sheets also facilitates the process of conducting an **item analysis.** After you administer a test, particularly after the first time you administer it, you may find it helpful to analyze the responses to each item by counting the number of people who marked each possible response, as in the sample table below:

SAMPLE ITEM ANALYSIS CHART

Number of responses: Item no. 3	A	B	C	D	E
upper quarter	0	0	5	0	0
middle half	1	2	4	0	3
lower quarter	0	1	1	1	2

The sample chart is for item number 3 of a test for which there were five possible responses, with C, the underlined response, keyed as the correct response. The terms in the left-hand column refer to those people who scored in the upper quarter of the total of twenty people who took the test, those who scored in the middle half, and those who scored in the lower quarter on the test. Note well that **the upper, middle, and lower portions of the twenty respondents are divided based upon their total test scores,** not on the basis of their answer to item number 3. The sample table reveals several important pieces of information. This item separated those who did well on the test from those who did not. All five of the people in the upper quarter correctly answered this item, those in the middle quarter answered it correctly about half the time, and those in the lower quarter almost never answered it correctly.

The fact that those in the lower quarter marked answers spread across the response categories indicates that those answers are viable alternatives. However, only one of the twenty people who took the test marked either response category A or response category D. Since none of the people in the lower quarter marked A, that response option may need revision; so might option D for the same reason. I would not undertake to revise those two response options given the data presented here, but would watch for similar patterns after administering this question to about seventy-five to one hundred people.

If you are trying to construct a more psychometrically sound test then you might undertake an item analysis which calculates correlations between individual item scores and the total score on the test, in an attempt to separate those items which match with the rest of the test from those which do not. Space does not permit a complete review of that kind of item analysis here. You can, however, find consultants who would be able and willing to assist you in such an undertaking when and if it becomes necessary.

6.5.3 CORRECTION FOR GUESSING

If you want to make a correction for guessing in your test scores, according to Thorndike and Hagan (1977, p. 250) the generally accepted formula is:

$$\text{Corrected score} = R - \frac{W}{n-1}$$

in which R is the number of questions answered correctly, W the number answered incorrectly, and n the number of response categories for the items. Thus, on a multiple-choice test with 4 possible responses to each item, the corrected score is R - (W/3). On such a test, if a candidate faced 40 questions, omitted 4, answered 24 correctly, and 12 incorrectly, the corrected score would be 24 - (12/3) = 24 - 4 = 20. Note that **no correction or penalty was assessed for the items which the candidate omitted.**

Obviously, candidates need to be told whether there is a correction for guessing in any circumstance in which their scores will have some effect upon their future, such as when their test scores will contribute to a final grade in the course or to their ability to qualify for certification. If there are **different item types in the same test then a different formula for correction against guessing** would have to be applied to items of each type. Consequently, you may find it easier to group true-false, multiple-choice, and matching items into subsets so that you can correct the score for each subset of items using the appropriate formula and then sum those corrected scores to get the total test score.

The correction for guessing helps to equalize the expected percentages of respondents who will correctly answer various item types. When I applied the appropriate formula to correct for guessing on each of the item types listed in the table of relative difficulties included earlier in this section, I got some revised figures for relative difficulties which are remarkably more similar than the earlier estimates:

Item type:	Relative Difficulty (%) Without correction for guessing	With correction for guessing
true-false, dichotomous-choice	85	70
multiple-choice		
3 response choices	77	66
4 response choices	74	65
5 response choices	70	62
cloze, sentence completion	50	--

In essence, the data in the above table show that when corrections for guessing are applied, then true-false items more closely resemble multiple-choice items in terms of their difficulty if only three response choices are given to the multiple-choice item. There are, however, still some appreciable differences between five response choice multiple-choice items and true-false or dichotomous-choice items.

6.5.4 INTERPRETING TEST SCORES

When you give a test as a pretest, you can gain valuable information about the knowledge which the participants in a program bring to your sessions. For that knowledge you trade the advantage of concealing the content of your test and the content of your sessions.

Based upon an item analysis of pretest scores, or a simple inspection of the test papers to see which test questions most respondents are missing on the pretest, **you can modify or completely shift the emphasis in your educational program from one topic area to another.** You may discover that although your needs assessment indicated that respondents did or did not need to hear about a particular subject, their pretest scores indicate the opposite.

If you give a test as a pretest and you hope to measure any changes in knowledge by giving a posttest, then you will need a parallel form of the instrument you used as a pretest to administer as the posttest. Obviously, if you administer the same instrument as both pretest and posttest, some gain might be expected through familiarity with the questions when the instrument is given as the posttest.

If you are using the test as either a pretest or posttest in an experiment, then you need to pay attention to the material regarding threats to

validity contained in **section eight.** One of the threats to the validity of an experiment is called **pretest sensitization;** that is, the subjects of your experiment become sensitized to what you expect them to know or to learn as a result of taking the pretest. That problem can turn into a blessing if you are not doing an experiment, because the pretest can pique the interest of the participants, alert them to certain material, and improve the chances that they will attend to crucial elements in the presentation.

6.6 USING TESTS CONSTRUCTED BY OTHERS

Sometimes a test that someone else developed or even published will suit evaluative needs. Since it is impossible to know in advance whether or not a search for someone else's test will bear useful fruit, the question immediately arises as to where to look and for what. When looking for tests, as in no other instance, it is imperative to read critically, search thoroughly, and adopt skeptically.

The end of this section will list several books which contain information about tests. In addition to those sources it is possible to find information about tests in issues of various journals. New tests are sometimes reviewed in journals, and authors of journal articles sometimes develop new tests for their research. In some instances the creation of the test itself, since it is a lengthy research project, will be the subject of an entire journal article. Also, those who publish tests advertise their work in journals.

Wherever you find the information about a test, you will want to know several things about the test before you adopt it. Any published test safe for use comes with a manual which includes all the information listed below. When no manual has been published by one of the major publishers, then the authors of the test, if they plan to distribute it, should provide a similar source of information.

You will of course be concerned with the validity and reliability of the instrument, both of which were discussed earlier. A manual or other source of information about a test should clearly specify how the validity and reliability were determined and report a reliability coefficient which the authors or publishers use to support their claims. The literature on the test should clearly reveal to you whether or not the instrument is appropriate for the group you will test. A very valid and reliable test

designed for high school seniors loses its validity and reliability if applied to the wrong population, such as junior high school students or senior nursing students.

If you think you have located a test which you can use, order a copy of the test and inspect it. You may find that the table of specifications used to construct the test you examine is not the same table of specifications as the one you need. Also, inspect the test for differences in language and reading level to ensure that the test is appropriate for your audience.

There are special problems with **attitude tests,** instruments designed to measure various kinds of attitudinal domains. When respondents mark an answer on an attitude test, they undergo a thinking process between the time the item is read and the answer is marked. The individual's attitude about the content of the item may play a small role, a major role, or no role in the process of choosing a response. Some people respond based upon what they believe is socially or professionally desirable as a response. Thus, when you use any sort of written attitude measure, exercise caution when drawing conclusions about the meaning of the results. Your ability to make conclusions about attitude changes will be particularly muddled when the people taking the test have a vested interest in the final outcome. Clearly, if you tell a group of new education volunteer trainees that their final approval to serve as volunteers will depend in part upon their scores on an attitude test then you can expect their answers to reflect a great deal of what they think you want them to say on a test, in addition to, or perhaps instead of, their true attitudes.

Buros, Oscar K. **Mental Measurement Yearbook** (8th ed.). Edison, N.J.: Gryphon, 1978.

Chun, K., Cobb, S., and French, J.R.P., Jr. **Measures of Psychological Assessment.** Ann Arbor, Mich.: Institute for Social Research, Univ. of Michigan, 1973.

Robinson, J.P., and Shaver, P.R. **Measures of Social Psychological Attitudes** (rev. ed.). Ann Arbor, Mich.: Institute for Social Research, Univ. of Michigan, 1975.

SELF-TEST ON SECTION SIX

You will probably gain the most benefit from this self-test if you write your answers on a separate sheet of paper and then compare them to the answers given on the pages following the self-test.

1. True or false:
   A. An IQ test is a norm referenced test.
   B. A final exam based upon the course objectives is a criterion referenced test.

2. Define validity and reliability.

3. What is content validity?

4. Suppose that you have written a new test on human reproductive anatomy and physiology. If you decide to examine the reliability of the test using the test-retest method, will you necessarily have to conduct the second administration of the test?

5. If you get a split-halves reliability correlation of .80, what is the maximum reliability estimate for your whole instrument? The Spearman-Brown Prophesy Formula is:

$$r_{11} = \frac{2(r_{\frac{1}{2}\frac{1}{2}})}{1 + r_{\frac{1}{2}\frac{1}{2}}}$$

6. The acceptable limit for minimum reliability of a questionnaire is __, but the minimum acceptable level for a knowledge test is____.

7. Can a test which is not reliable be valid?

8. What is a table of specifications?

9. Critique each of the following test items:

   A. An IUD should never be inserted into a woman who has had more than one abortion or who has expelled two IUDs before. (True or false.)

   B. Married couples sometimes have difficulties with the diaphragm. Which of the following reasons is a common concern about the diaphragm?
      a. It requires no cooperation from the male.
      b. It requires the woman to touch her genitals.
      c. It has relatively low effectiveness.
      d. All of the above.

   C. Match the following items with their use effectiveness:

1. _____douching	A. 60%
2. _____withdrawl	B. 75%
3. _____the pill	C. 96%
4. _____the IUD	D. 93-96%

10. All of the following are advantages of having prepared answer sheets EXCEPT:
    A. they make scoring easier and quicker.
    B. they enable you to reuse the test booklets.
    C. they increase the reliability of the test.
    D. they increase the validity of the test.
    E. All of the above are advantages of prepared answer sheets.

11. Define item analysis. Why would you want to do one?

12. What is a formula to correct for guessing intended to do in terms of relative item difficulty?

ANSWERS TO SELF-TEST ON SECTION SIX

1. A. True
   B. True

2. A test has validity if it measures what it purports to measure. Psychometrists recognize at least four kinds of validity: content, construct, concurrent, and predictive.

   Reliability refers to the ability of an instrument to produce the same result when administered on different occasions to the same individual or group.

3. Content validity exists when a test measures a representative sample of the total possible universe of material which it purports to measure.

4. No. If you administer it once and calculate a split-halves correlation which yields a sufficiently large Spearman-Brown Prophesy Formula estimate of reliability, you do not need to administer the retest, assuming that you have met the criteria for split-halves estimates in 6.2.2.

5. r_{11} = .84. This example shows that the closer the split-halves reliability gets to 1.00, the smaller the difference between r_{11} and $r_{\frac{1}{2}\frac{1}{2}}$.

6. About .80; .90, but do not worry much about .85 unless you plan to use the test for experimental research, in which case you may have a problem if your experimental results are not significant. If you use a test to measure differences between experimental groups and you get a borderline difference when you apply inferential statistics (9.3) you cannot know for sure if the experimental groups do not differ or if the reliability of the test masked a truly significant difference.

7. No. A test which cannot reliably measure anything cannot be a valid measure of anything.

8. A table which indicates the percentage of test items which will be devoted to various topics based upon some criteria such as the opinion of a panel of experts or the percentages of time spent on various topics in a course or other educational program.

9. A. The word "never" nearly always indicates that the answer is false. There are actually two items in this question, one about women with histories of multiple abortions and one about women with histories of IUD expulsions.

   B. The first sentence is irrelevant and misleading; some women who use diaphragms are not married. Answer a is a matter of opinion. Response alternative c is poor because "relatively ineffective" implies a comparison, but not to what. Response d is poorly utilizied in this case.

   C. Even a slight familiarity with contraception enables a candidate to know that A and B must match with 1 and 2 and that C and D must match with 3 and 4. Answers C and D do not differ signicantly, making a certain answer impossible to identify. There should be a fifth response option.

10. D. If you reasoned that a prepared answer sheet increases validity by increasing reliability, you were incorrect. Increasing reliability does not ensure increased validity; although reliability is a necessary condition for a test to be valid, it is not a sufficient condition.

11. Item analysis involves an examination of the response patterns to a test in order to identify items which separate those who know the material from those who do not. Poor items reduce the validity and reliability of a test. Thus, you would want to do an item analysis as a matter of course for a test you plan to reuse several times, or as a method of identifying poor items on an unreliable test.

12. The correction for guessing more nearly equates the difficulty of the various item types, thus giving a test constructor more leeway in using item types on a test. It also tends to improve the reliability of a test by helping to eliminate random chance as a factor in the final scores for individuals.

section seven:
SURVEYS

After reading this section and learning the material in it you will be able to:

1. Describe the following kinds of surveys:

    a. mass
    b. case studies
    c. correlational
    d. retrospective
    e. longitudinal
    f. cross sectional
    g. comparative

2. Indentify a seven-step sequence to follow when designing, conducting, and analyzing the results of a survey.

3. Describe the advantages and disadvantages of each of the following methods of contacting a survey sample:

    a. mail surveys
    b. telephone interviews
    c. canvassing
    d. interviewing

4. Define and state the importance of representativeness in sampling.

5. Define and differentiate among:

    a. random sampling
    b. stratified random sampling
    c. deliberate sampling
    d. cluster sampling
    e. sequential sampling

If you are familiar with the material in this section, you may want to turn to **section eight.** If you are not sure, or if it has been a while since you studied surveys and sampling, you might want to turn to the self-test at the end of this section.

section seven: SURVEYS

"Curiosity is one of the permanent and certain characteristics of a vigorous mind."

--Samuel Johnson,
The Rambler,
March 12, 1751

Survey research methods will enable you to gather information from groups of people in the most efficient and effective way possible. A survey is appropriate when you need information about the characteristics of a population or group, whether it is very small or very large, but you do not have access to the entire group or you do not have access to the group at one place and time.

You may find surveys useful when you conduct needs assessments. Some family planning agencies conduct marketing surveys--surveys of samples from the communities they serve--using a portion of a state, an entire state, a region, or even the nation in order to gather data about the public's opinion of family planning, a particular agency, or issues related to reproductive freedom. Surveys may assist you in performing evaluative follow-up for professionals who invited you to do educational programs for the people they serve.

In some cases you may be able to offer every member of a group an opportunity to respond to your survey. However, surveys are most useful and most often conducted in situations where the entire population cannot be directly contacted for logistical reasons or because the size of the population would make an effort to contact them all prohibitively expensive. When you cannot contact the entire population, you need to select a **sample**, a subgroup of the population.

The first subsection below descibes the various types of surveys so that you can choose the survey method that best meets your needs. Next come the steps to follow when conducting a survey. The final subsection describes various methods of sampling from a population.

7.1 TYPES OF SURVEYS

Although you may never need to use some of the kinds of surveys described below, an awareness of all of your options will increase the likelihood that you make maximal use of the survey approach.

Mass surveys. The United States census is the most famous example of a mass survey. While the census provides relatively little information about each individual, the results provide volumes of information and data about the demographic characteristics of the US population. Mass surveys are very expensive to conduct because they involve an effort to contact every member of a given population. I have difficulty envisioning a situation in which a family planning educator would attempt to conduct a mass survey of a group larger than a few hundred people, mainly because of the time and expense involved.

Case studies. The case-study approach represents the opposite end of the spectrum from the mass survey. With a case study you select a relatively small group of individuals from a larger population and survey their attitudes, behaviors, experiences, lifestyles, or whatever, in great depth. While the limited size of a case study survey sample usually prohibits making generalizations about the entire population, case studies can provide invaluable information in the form of questions to be answered by other kinds of surveys or by experimental research.

Correlational surveys. Education is a relatively new discipline, and much of the background data which educators need is yet to be gathered. Researchers frequently conduct surveys for the purpose of discovering relationships between or among groups of measures or variables. Once they have collected their data, they calculate correlation coefficients (see 9.3.2) in an attempt to discover relationships among important characteristics or behaviors. When correlations do exist, researchers can sometimes use the relationships among their data to make predictions about the future. Those predictions are made using an advanced statistical technique called **regression analysis** or the more sophisticated technique called **multiple regression analysis.** (See 9.3.)

Retrospective surveys. If you decide to ask people about their experiences, behaviors, or characteristics during times past, you are conducting a retrospective survey. You could say that any survey is retrospective; the distinction is that a retrospective survey is one which attempts to understand the present based upon information about the past.

Longitudinal surveys. One of the more frequently asked kind of question, and least frequently answered, concerns whether or not an educational program had an effect which lasted over some period of time, say six

months, a year, or more. Of course, the only way to ascertain whether or not a program had an effect which lasted is to contact the participants at several times after the end of the program to see what effect the program had upon them, their lives, their professional performance, or the like. Research surveys aimed at doing this are called longitudinal surveys. These kinds of surveys are extremely expensive and time consuming, particularly because they require extensive record-keeping and frequent updating of the addresses and phone numbers of the participants. In the ideal situation, family planning educators would conduct longitudinal surveys following many, if not most, educational programs.

Cross sectional surveys. When a longitudinal survey cannot be conducted, some researchers use an approach which requires that they simultaneously survey people at various stages of their development or at various lengths of time following a specific kind of educational program. Instead of following one group of individuals as in a longitudinal survey, the researchers find several groups of people who are similar with respect to important characteristics and survey them all at once, hoping to gather data which reveal the same kinds of information as a longitudinal survey. But since a cross sectional survey does not follow the same group of individuals over a period of time, the results are muddled by the differences among the groups chosen for the cross sectional survey.

For instance, suppose you want to know what kind of long-term effects your new-patient education has upon your patients. If you choose to follow one group of patients for the year following their first visit to your clinic, you are conducting a longitudinal survey. However, if time and money are short, you might choose to conduct a survey among three different samples of patients: one sample who just finished your new-patient instruction; another sample who came for the first time six months ago; and a third group who came for the first time a year ago. If you choose the latter approach you are conducting a cross sectional survey.

Comparative surveys. When a researcher conducts an experiment, groups of individuals are assigned to different kinds of educational situations and the differences between the various situations are compared in terms of outcome, impact, or both. A comparative survey accomplishes the same end, but instead of creating the experimental situation, the researcher uses already existing educational programs and conducts a survey for the purpose

of comparing the two approaches, methods, or whatever. Comparative surveys are frequently used for evaluative purposes, particularly in public school education settings.

Combining types of surveys. Evaluators and researchers frequently combine two or three of the types of surveys described above in order to gather more information than any one type would provide in a given situation. For example, some people conduct case-study surveys in a longitudinal manner. Another researcher might conduct a comparative survey using a combination of case-study and cross sectional approaches, thus using three different techniques for one research or evaluation project. Combining types of surveys is frequently done and perfectly legitimate.

7.2 STEPS IN CONDUCTING A SURVEY

The following seven steps will suffice when you want to conduct a survey of a relatively small sample or population, say a group of one hundred or less. If you need to conduct a larger or more involved survey, then you might need to hire consultants, such as a public relations or opinion surveying firm, or you might enlist the assistance of volunteers who have expertise in conducting surveys. In addition to the cross references which I will make to other sections of this manual, you may find some of the references in the **bibliography** useful.

Step 1. Define the purpose of the survey in terms of behavioral objectives. What do you want your survey to tell you? About whom? Why? If you had the data from your survey in front of you right now, what would the data tell you that you don't already know?

At this point you need to consider the possibility that someone else has already gathered the data which you need **(see section four).** You may also decide, at this point or further down this list of steps, that a survey is not the best approach for learning what you need to know. A survey is appropriate when you can specify the kinds of information you want to collect and the groups of individuals from whom you want to collect it. In addition, surveys are appropriate when you can gather the information you need by conversing with your respondents or by having them complete a questionnaire, whether they mark the questionnaire themselves or answer the questions during an interview. If your information can come only from a process of having people attend some sort of an educational session, then you may need to consider an experiment **(see section eight).**

Step 2. **Select the type of survey which best suits your needs.** From the list of survey types in the preceding section choose the kind of survey or combination of types which will best provide you with the data or other information which you need. The answer to this question might have to wait until you have performed the tasks outlined in the next step, but you usually need to define the type of survey you want to conduct before you can select your population.

Step 3. **Define and identify the target and accessible populations.** In any survey there is a group of individuals, which may vary in size from five to fifty million, at whom you are aiming your survey. This group is called the **target population,** or the population **to which you wish to generalize the results** of your survey. However, only infrequently will you be able to conduct a survey among a sample of your target population, and even more infrequently will you be able to survey the entire target population. If your target population is small, then there is no reason you cannot survey the entire group. When you have to settle for a subgroup from your target population, **the group available to survey** is called your **accessible population.** If you hope to generalize your results from the accessible population to the target population, then you need to do everything possible to ensure that your accessible population is representative of the target population in the sense that **representativeness** is defined in 7.3 below.

You may need survey results for all junior and senior high school health teachers in your county. But you may not have access to that entire population; some of the school districts in your county may not be willing to give you lists of their names; some of the school districts' records may not be up-to-date; some of the school principals may not want their teachers to complete survey questionnaires. Ultimately, when you analyze your results and draw conclusions from them, you will have to deal with the question of how well the people who responded to your survey represent your accessible population and, in turn, how well that group who responded to your survey represents the target population. Therefore you need to make sure that your accessible population represents the target population as well as possible, in order to avoid problems of interpretation when you examine your data.

Step 4. **Choose the method you will use to contact your sample of respondents.** You may occasionally be able to contact your entire accessible population or your entire sample in one place at one time. Most of the time, however, you will have four methods from which to choose for contacting the members of your sample.

Mail surveys consume relatively little time and do not cost as much per respondent as some other approaches. However, mailing a questionnaire to a sample of respondents will typically yield no more than a 35 to 40 percent response rate; a 50 percent response rate is considered good and 60 percent is considered outstanding. Thus, when you analyze your results you have the problem of trying to decide how well the sample who responded represents the total sample to whom you sent the questionnaires.

You can increase the response rate by several different methods. If you include a self-addressed stamped envelope with the questionnaire, your response rate will increase by between 5 and 10 percent. Offering people rewards or prizes for returning the questionnaire seems to have relatively little effect upon the response rate because most people do not trust the researcher enough to believe that the reward will be forthcoming when the questionnaire is returned. Some surveyors keep a record of who has responded to a mail survey and then send a copy of the questionnaire to people who have not responded after a period of two to three weeks. Others send postcards to every respondent in the sample asking them to return the questionnaire if they have not already done so. Follow-up postcards may increase the response rate by 5 to 10 percent if they are sent relatively soon, say within seven to ten days after the initial mailing of the survey questionnaires. Much longer than seven to ten days greatly increases the chance that the respondents have discarded the original questionnaire without completing it; hence you may need to consider a follow-up mailing of another blank questionnaire to the sample.

Telephone interviews will greatly increase the percentage of respondents from a sample over the percentage you would get from a mail questionnaire, but the amount of time and effort involved in administering a questionnaire to a sample of respondents through telephone calls can become staggering very quickly. The most rational method for conducting a telephone interview survey involves constructing a questionnaire which everyone will use when interviewing respondents. If the questionnaire takes longer

than five to ten minutes to complete, the number of respondents will drop dramatically. Telephone interviews also limit the kinds of questions which you can ask to ones which require yes-or-no answers or only a limited number of possible responses. If you decide to conduct telephone interviews, you need to construct a questionnaire which is as explicit and easy to read as possible, and then train every person who will be doing the telephone interviews to make sure they all ask the same questions in the same manner and record their answers in the same consistent fashion.

You can combine telephone interviews with mail surveys in two different ways. Some people phone those people who did not respond to a mail questionnaire to attempt to get their responses on the telephone. You might also want to consider conducting a mail survey and a telephone survey among different subgroups from the same sample, hoping to get more complete information from a small telephone interview sample than you get from the larger mailed questionnaire sample.

Canvassing for interviews involves going door-to-door to conduct face-to-face interviews with your sample, in their homes or in their places of work. The effort required is similar to that which door-to-door salespeople have to expend to meet sales contacts: people have to be visited more than once because they are not available; people refuse to respond to the survey even if you can find them, and so forth. However, major polling organizations such as Gallup and Harris use this method for their famous polls because it greatly increases the chance of obtaining an accurate result. Of course, people like Gallup and Harris have an awful lot more money than most family planning educators have for conducting surveys.

Passers-by, such as people who happen to enter a shopping mall or attend a convention, can sometimes be interviewed quickly (in less than three to five minutes). However, this approach often causes problems when you try to generalize your results from the accessible population to the target population. In other instances this is not such a serious problem. For example, if you enter a shopping mall and interview people who happen to pass by to solicit their opinion of your agency, then your accessible population includes only people who go to shopping malls, a group which typically excludes low-income, and consequently minority, populations.

On the other hand, if your target population is accessible to interviews, as members of the population happen to pass by, then there is no

such problem. For instance, you might conduct a valid survey of your patients by interviewing patients who happen to be sitting in your waiting room at various times of the day over the course of a week. If your target population is your patients, then interviewing people sitting in the waiting room makes it easy to generalize from your accessible population to the target population.

Step 5. **Prepare your survey questionnaire or interview form.** The section on writing questionnaires, **section five** of this manual, describes the procedures for writing a questionnaire or constructing an interview form. See especially the last portion of **section five,** which covers pretesting of questionnaires.

Step 6. **Decide how you will select your sample and how large the sample will be, unless you plan to survey the entire accessible population.** Subsection 7.3 discusses sampling techniques.

Step 7. **Conduct your survey and analyze your results.** The material in **section nine** should be of assistance in compiling and analyzing the results of your survey.

If you have taken a good deal of time and effort with your survey, or if you need to document the results, you will want to write a report describing the process, the data, and your conclusions. Minimally, the report should compare the data to the questions which you asked in the form of behavioral objectives in **step 1.** Your survey may answer the questions which you set out to answer, but it may also raise more questions than it answers. In order to minimize the problem of raising more questions than answers, you might sit down with your complete survey questionnaire and imagine various kinds of response patterns and then attempt to interpret those response patterns. That kind of forethought may lead you to revise, edit, or expand your questionnaire. Once again, refer to the last portion of **section five.**

7.3 SAMPLING

Two considerations always conflict when making choices about samples, their selection, and their size. First, the sample must be **representative** of the accessible population, and should allow one to extrapolate from the sample to the accessible population and then to the target population. Ideally, the sample is representative of the target population itself. Second, the process of selecting and surveying the sample should be done as

efficiently and economically as possible in order to avoid waste of time and other resources.

Representativeness should not be sacrificed in order to avoid spending time and effort on the survey. Otherwise the data collected in the survey will be seriously compromised. The factors which dictate whether or not a sample is representative vary greatly from one situation to another. Frequently, demographic characteristics of respondents, such as age, sex, educational level, socioeconomic status, location, and the like, are important. Sometimes prior reserarch, whether experimental or survey in nature, will give clues as to those characteristics of a sample which will be important when representativeness is examined at the time the data are analyzed. Frequently, and particularly in educational fields, guesses, hunches, and suspicions are all one has upon which to rely. The representativeness question ultimately concerns whether or not the sample which responds to a survey matches a target population with respect to those characteristics which we know, or suspect, will have an effect upon the responses to the survey.

There are two ways to increase the representativeness of a sample. The first is to choose as large a sample as possible, thus hoping to dilute individual differences among the respondents. The other sampling decision concerns the method of choosing the sample.

As a general rule, it is wise to choose the largest possible sample within the limitations upon your time and other resources. However, in those rare instances when you can use a sample as large as fifty percent of your accessible population, you will be on safe ground if you limit your sample size to that fifty percent. Of course, if you can survey the entire accessible population, you need not worry about the ability of the sample to represent the accessible population, unless your response rate is very low.

Random sampling. You have chosen a random sample when you select individuals from the accessible population for inclusion in the sample in such a manner that **every member of the accessible population has an equal chance of appearing in the sample.** Beginners frequently confuse random sampling with **systematic sampling.** Suppose you have a list of two hundred health teachers and you want to select a sample of fifty. If you arrange the list in alphabetical order and select every fourth name, you are con-

ducting a systematic sampling process. That process is not random because once you choose the first name from the list and then begin to count off every fourth name, you have a predetermined sequence. If you assign a sequential number to the names on the list and then consult a table of random numbers (typically found in elementary statistics books) and select those people whose numbers happen to appear in the table, then you are selecting a random sample.

Random sampling is the most common technique for ensuring representativeness. The theory behind random sampling is that the individual differences which may affect the outcome of the survey will be balanced. Ideally, the random sample will contain neither a lack nor an abundance of individuals exhibiting those characteristics which may influence the results. Only rarely will you be able to identify even a portion of the list of characteristics which affect the results of the survey. The remaining unknown characteristics which will affect your results are assumed to be balanced by the random selection process.

Stratified random sampling. When you can specify some of the characteristics of your **target** population which are likely to have an effect on the results, then you can select a **random sample containing proportionate representation** of those characteristics of the target population. Suppose that you wish to select a sample of 100 people from an accessible population of 1,000 people. Suppose, further, that you know that race and sex will affect the results of the survey you are going to conduct. Your **target** population contains 10 percent blacks and 90 percent whites with 60 percent women and 40 percent men. Assume that both racial groups have the same proportions of men and women as the total sample, 60 percent men and 40 percent women. Then, in the parlance of sampling, you have four **strata:** black males, black females, white males, white females. You would want 10 percent of your sample of 100, or 10 people, to be black and the other 90 to be white. Within your subgroup of 10 blacks you would want 60 percent females, or 6 black women, and 40 percent male, or 4 black males. Similarly, you would want 60 percent of your white respondents, or 54, to be women and the remaining 40 percent, 36, to be white males.

As you read the last paragraph, it probably occurred to you that you would have very small representation in your sample for the black male and black female strata. However, if your accessible population varies greatly

from the characteristics of your target population with respect to race and sex, then a random sample might over-represent groups which could affect the results of your survey. Suppose your accessible population contains almost no blacks, or almost no women. Then a random sample from the accessible population is very unlikely to generalize to the target population.

There is another option which you could use to make your results more representative, an option which you would apply at the time you analyzed your data. Suppose, for example, you are facing the same situation in the example above but you decide to simply select a straightforward random sample from your accessible population. Suppose you have administered your survey questionnaire and gathered your results. If you discover that your sample contains disproportionate representation from one racial group or the other, or one of two sexes, then you can analyze your results by race and sex and then adjust your response pattern by calculating proportions to elimate the discrepancies between your sample and the target population.

Using either random sampling or stratified random sampling eliminates biases in the results and challenges to your conclusions on the basis that your sample was not representative. There are, however, instances in which you may want to use other sampling techniques.

Deliberate sampling. Sometimes you may decide to deliberately include individuals or whole groups in your sample because their responses are of particular interest or would be of crucial value in interpreting your results. This deliberate sampling approach is clearly not random sampling, but may be justified by the objectives which you have stated at the outset of your survey process. You should only choose deliberate sampling when you do not care whether or not your results from those portions of your sample who were deliberately selected will generalize to the target population.

Cluster sampling. The realities of sample selection and the logistics of survey techniques sometimes dictate using entire groups of individuals as portions of a sample. For instance, you may have access to groups of teenagers only by distributing questionnaires to them in high school classrooms, Sunday school classes, or at meetings of youth groups. Thus, clusters of individuals are included in the sample. Every person who happens to be present when you distribute your questionnaire is automatically included in the sample.

Cluster sampling is obviously not random, but you can overcome this problem by administering your survey to a group several times larger than you need for your sample, defining the people who were selected in the clusters as your accessible population, and then randomly selecting questionnaires for inclusion in your survey sample. You might conduct that sampling process by using either random or stratified random sampling techniques. This approach will permit you to generalize your results to your target population only if you can demonstrate that the clusters of individuals you chose are representative of the target population. Then you can claim that the questionnaires which you've randomly selected from among those gathered during the cluster survey adequately represent all those who completed a questionnaire. You can generalize to the target population by logically demonstrating that your clusters, in fact, are representative of the target group.

Sequential sampling. There is a sampling technique which will help deal with both of the concerns I mentioned at the beginning of this subsection: representativeness and efficiency. Suppose that you have access to a sample of 1,000 people but you have no indication as to which characteristics of those individuals may affect your survey results. Thus you do not have the option of using stratified random sampling, and you face the question of how large a sample you need to select by random methods.

Sequential sampling would involve selecting a sample of about 50 from your accessible population of 1,000, and a companion sample of 50. Then you administer your questionnaire to both samples of 50 and compare the results of administering the questionnaire to the two samples. If there is very little variation in response pattern between the two samples, then you are reasonably certain that additional responses would be unlikely to affect the response pattern. At that point you would combine your two samples of 50 into one large group of 100, compile their responses, and report your results and conclusions.

If, however, there were discrepancies between the results of the first two samples of 50, then you would collect data on a third sample of 50 and proceed as follows. You would combine the results of the third sampling of 50 with the first two, making a total sample of 150, and see if the results from that sample of 150 were significantly different from the results of the first two combined samples of 100. If not, then you would compile and

report the results. If, however, the third sample's results differed from the results of the first 100, you would repeat the process again, and again if necessary, until no remarkable changes occurred.

The risk that you run if you choose to conduct sequential sampling is that you may have to repeat the process several times before you see no remarkable change in the results. However, sequential sampling typically requires no more than two or three samples because the theory of random sampling helps to ensure that the results will remain fairly constant. What you guard against with sequential sampling is the unlikely chance that your sample was so small in the first place that it caused unrepresentative results.

7.4 OTHER PEOPLE'S QUESTIONNAIRES

Sometimes other people's questionnaires may be adapted to your situation. The guidance offered at the end of **section six** with respect to locating other people's instruments and evaluating their applicability to your needs applies here too.

The excellent work done by several individuals deserves mention.

Mathtech (Mathematica, Inc., 4630 Montgomery Ave., Bethesda, Md. 20014) performed an extensive evaluative review of sex education for the Center for Disease Control, resulting in a multivolume work which includes several questionnaires: US Department of Health, Education, and Welfare, Public Health Service. **An Analysis of Sex Education Programs and Evaluation Methods.** Atlanta: Center for Disease Control, Bureau of Health Education, 1979. (Report Number CDC-2021-79-DK-FR.)

Dr. Guy Parcel (Univ. of Texas Medical Branch, 1202 Market, Galveston, Tex. 77550) has published findings from the development and application of several instruments which may apply to your work. Consult his name in the various indexing services and abstracting publications.

Planned Parenthood of Santa Cruz County (Calif.) also has developed several evaluation questionnaires for teacher training and other applications.

Perhaps the best-researched sexual knowledge and attitude questionnaire or inventory is **The Sex Knowledge and Attitude Test** (Lief & Reid, 1975).

SELF-TEST ON SECTION SEVEN

You will probably gain the maximum benefit from this self-test if you write your answers on a separate sheet of paper and then compare them with the answers on the pages following the self-test.

1. What kind of survey do you conduct if you:

   A. Attempt to contact every health education teacher in your county.

   B. Gather data and then calculate measures of association (see 9.3.2) for various sets of data.

   C. Interview three or four new patients in depth about their experiences in your clinic.

   D. Interview a group of former board members to find out what kinds of radio PSAs your agency used ten years ago.

   E. Send questionnaires through the mail to trainees one month, three months, six months, and one year after they complete a training course.

   F. Send questionnaires to three samples of new patients simultaneously: one sample who visited for the first time in the past month, one sample who were at your clinic six months ago, and one comprised of patients who came for the first time more than a year ago.

2. What is the difference between a target and an accessible population? Why is this of importance to an educator conducting a survey?

3. If you send out 100 questionnaires to a sample through the mail, you might expect to get back about_____responses and you would have reached the usual maximum response rate if you got back _____.

4. Suppose you have a population of 1,000 high school teachers whom you want to survey. You need to select a sample of 100 from the alphabetical list of 1,000. What kind of sampling do you use if you select your sample with each of the following methods?

   A. You put their names in a hat and draw 100 names after thoroughly mixing the slips of paper in the hat.

   B. You close your eyes and point to one of the first ten names on the list and then choose every tenth name down the list.

   C. You divide the list into eight groups: male teachers and female teachers from each of four school districts. Then, you use a table of random numbers to select the same proportion of men and women from each school district as the proportions that appear in the entire population of 1,000.

D. Choose those teachers whose classes you have visited and pick the remainder by drawing names from a hat.

5. Suppose that you have the same population as in question 4. What kind of sampling do you use if you select two samples of 30, conduct your survey, compare the results of the two samples of 30, and then repeat your process by sampling another 30 if the first two samples show different response patterns to your survey?

6. Define representativeness as it relates to sampling.

ANSWERS TO SELF-TEST on SECTION SEVEN

1. A. mass
   B. correlational
   C. case study
   D. retrospective
   E. longitudinal
   F. cross sectional

2. The target population is that group to whom you hope to generalize your results. The accessible populaton is the group you have available for drawing a sample for surveying. If the accessible population does not adequately represent your target population, then in spite of the fact that you may have used sound sampling procedures, you will have difficulties when you want to generalize the results of your survey to the target population from the sample.

3. 30-35; 50. You can increase the response rate by sending reminder postcards or another copy of the questionnaire to the entire sample or to those who have not responded after about seven to ten days.

4. A. random
   B. systematic (which is NOT random)
   C. stratified random
   D. a combination of deliberate and random sampling

5. Sequential sample.

6. Representativeness means that the sample matches the **target** population with respect to those characteristics which could affect the results of the survey. Representativeness should always supersede considerations of time and effort when choosing a size for your sample.

section eight:
EXPERIMENTAL RESEARCH

After reading this section and learning the material in it you will:

1. Have guidelines for determining circumstances under which you might want to conduct experimental research.

2. Be able to define and state the importance of the following threats to the **internal** validity of experiments:

   a. history
   b. maturation
   c. testing
   d. instrumentation
   e. statistical regression
   f. selection of subjects
   g. mortality

3. Be able to define and state the importance of the following threats to the **external** validity of experiments:

   a. accessible vs. target population
   b. interaction of treatments and personological variables
   c. description of independent variables
   d. multiple treatment interferences
   e. The Hawthorne Effect
   f. novelty and disruption
   g. posttest sensitization
   h. measurement of the dependent variables
   i. experimenter bias

4. Be able to analyze and judge the validity of your own and other people's experimental designs and results.

5. Describe and, by referring to other parts of this manual, use behavioral observation as a data gathering technique.

6. Be able to converse with consultants about experimental research, with sufficient background to enable you to make maximal use of the consultant's time.

NOTA BENE: You might be wise to read and learn the material in **section nine: ANALYZING DATA** before you attempt to conduct an experiment, in order to ensure that you take into account the statistical considerations necessary to ensure that your experimental data will yield what you want.

A self-test appears at the end of this section.

section eight: EXPERIMENTAL RESEARCH

"Observation is a passive science; experimentation, an active science."

--Claude Bernard, 1865

Situations may arise in which you will want to compare two or more approaches to the same educational situation by conducting an experiment. You conduct an experiment when you create educational situations under carefully planned conditions and then measure the impact, outcome, or both (see 3.1) of different educational contents or processes.

Experimental research requires much planning, time, effort, and other resources. Therefore, limit the number of situations in which you decide to conduct experimental research to those questions of basic importance or extreme significance in your overall planning and evaluation process. Shoddy experimental research, done without taking the necessary pains to ensure that the results will be meaningful, simply wastes time and effort.

You will probably be able to conduct relatively short and simple experiments if you follow the guidelines presented in this section. However, if you plan an extended research project, you will probably find it helpful to have the advice of a consultant who is well versed and experienced in educational research situations.

Before you attempt to conduct any experimental research, you would be wise to review the literature on family planning and sexuality education, to make certain that nobody else has already done your work for you. Even if no one has conducted an experiment which is similar enough that your planned experiment becomes unnecessary, you may find valuable information which helps to guide the planning of your research project. A carefully conducted experimental research design based upon a solid understanding of prior research may enable you not only to answer the questions which are of interest to you, but also to write a research paper which you can publish and share with others.

Rather than present a general guide to experimental research, I will define a specific research problem and then design an experiment which answers the questions raised by the research problem. The remainder of this section will be an examination of the research design with respect to the principles of experimental research in education.

8.1 SAMPLE RESEARCH PROBLEMS AND DESIGNS

Suppose you are an education director who is responsible for beginning to educate all new patients who enter your family planning clinic. While the process of patient education never really ends and is the joint responsibility of all staff and volunteers--whether educators, clinicians, counselors, cashiers, or whatever--you must fulfill the responsibility of beginning the process and imparting a basic body of critical information. This responsibility requires that you offer education to approximately two hundred new patients monthly. Individual sessions with each of those two hundred patients would require that two full-time employees devote their entire working day to educating new patients every day. Consequently, you have decided that the only feasible method for educating new patients is to do so in groups.

You have two methods of educating new patients in group sessions. The first method, call it the lecture method, consists of a one-hour-and-fifteen-minute lecture including a fifteen-minute film on pelvic examinations. During the lecture, the instructor describes your clinic procedures for new patients, a step-by-step sequence of events which a patient experiences during a visit to your clinic, and the various methods of contraception which you have to offer. The second approach, the audio-visual method, presents the same information to the new patients in the form of a 35 mm slide-and-tape cassette program projected onto a screen large enough to be seen from anywhere in a lecture room, or onto a small table-top unit for an individual patient.

Each of these two methods has its obvious advantages and disadvantages. The lecture method enables more direct human contact between the instructor and the new patients, and makes it easier for patients to ask questions and hear their answers as the questions arise. However, since you have two hundred new patients to educate each month, this means that someone or several people will have to repeat the same lecture fifteen or twenty times monthly. Repeating the same lecture will not only cause burn-out, but will also lead to discrepancies in the quality and content of instruction from one class to another. The less personal audio-visual approach will reduce burn-out of the educators, provide accurate and consistent instruction, and accommodate those patients who cannot attend the regularly scheduled classes, without requiring a separate individual lecture.

A careful analysis of the cost-effectiveness and other advantages and disadvantages of the two approaches would still leave important questions about which method to choose. Which approach imparts the most factual information to patients? Which approach is more satisfactory to the new patients, thus increasing the chances that they will return as continuing patients? Which method most simplifies the process of obtaining a method of contraception when the patients have left the class and gone to their clinic appointments? Which of the two methods most effectively accomplished the "outcome" objectives for your new-patient instruction sessions?

This is an excellent example of a situation in which an experiment to compare the two methods of instruction could provide important information to aid your decision-making process. Stating the research problem is the first step. The questions contained in the last paragraph are an excellent statement of the research problem itself.

After examining those research questions, we find that there are three different problems: one concerning the amount of knowledge which the two methods impart, one concerning patient satisfaction with the method, and one concerning which method best prepares patients for a trip to the clinic. Therefore, three different measures of the outcome of the two different kinds of instruction would be required in order to compare the two methods on all three criteria.

The experiment might be designed in the following manner. Over the course of a month, offer instruction using each of the two methods to ten different classes of the new patients. Alternate the two methods using each on all of the various days of the week and at each of the various hours of the day when new-patient classes are offered. This approach would eliminate any differences caused by variations in the kinds of new patients who attend at different hours of the day or on different days of the week.

Following each of those twenty new-patient sessions, administer two instruments to each participant. The first instrument would be a test constructed to measure the new patients' knowledge of the information covered in the new-patient sessions. Construct the test based upon the guidelines presented in **section six,** using written behavioral "outcome" objectives for new-patient information as a basis for the table of specifications (see 6.3.1). The second instrument would be a questionnaire designed to measure whether or not the patients felt satisfied with the new-

patient instruction which they received. Constructing questionnaires is covered in **section five.** One would administer the same test and the same questionnaire to every new patient who attended all twenty sessions, whether they received instruction through the lecture method or the audio-visual approach.

Then I would use an approach called **behavioral observation** to compare the two instructional methods in terms of how well they prepare patients for their visit to the clinic. Behavior observation is a technique for gathering data about human interactions which attempts to reduce the subjectivity involved by as much as possible. First, one would write a list of behaviors which signify that a patient is experiencing greater or lesser difficulty with their visit to the clinic, per 2.2.4.

I would probably interview some clinic workers to ascertain what kinds of things they hear, see, or otherwise observe about new patients, things which make the patient's visit to the clinic easier or more difficult. Then I would construct a questionnaire which specified those behaviors which clearly indicate, or seem to indicate, whether a patient will have an easy trip through the clinic.

For instance, suppose that the interviews with current staff members revealed that patients who knew what to expect during the pelvic and breast exam seem to experience less pain and anxiety, and to complete the exam more quickly than those who do not. Suppose also that patients who can quickly state their first and second choices for a method of birth control during the pre-exam counseling session require less time during the session than those who cannot. Those criteria translate into behavioral statements like the following ones:

The patient described the various parts of a pelvic and breast exam when asked to do so by the doctor or nurse practitioner.

The patient could specify her first and second choices for a method of contraception within ten seconds of being asked to do so during her pre-exam interview.

If the interviews with clinic staff members indicated that some patients have no problem with pain or discomfort during their pelvic exam, while others seem to experience some discomfort, I might add the following behavioral statement to the questionnaire:

When the speculum was inserted during her pelvic exam, the patient did not exhibit verbal or nonverbal expressions that she felt discomfort.

Next, one would arrange the resulting statements in chronological order based upon the sequence of events during a visit to the clinic by a patient. Then, during the experimental period, the behavioral observation questionnaire would enable you to gather data with respect to all patients given new-patient instruction by either method.

There would be two ways to gather that observational data: a small group of trained observers, perhaps some educators and some clinic workers, could follow each new patient through the clinic process, observing the process and completing the behavioral observation questionnaire; or each person who comes in contact with the patient could complete the section of the questionnaire relevant to his or her portion of the patient's clinic visit. Either approach would require that the list of behavioral statements permit the people completing the questionnaire to easily determine whether or not each item applied to a particular patient.

Scoring the behavioral observation questionnaires would involve simply counting the number of behavioral statements which applied to each patient's visit. Thus, a high score would indicate that the patient had an easier visit than a patient who had a low score . Obviously, this would mean wording each of the behavioral statements in such a way that a "yes" or check-mark beside the statement would indicate a more positive experience than a "no" or no check-mark.

At the end of the experiment, three outcome measures would be used to compare the two instructional methods: a direct comparison of the two groups of new patients' test scores on the knowledge instrument, using the inferential statistical techniques discussed in 9.3.3, would reveal differences in their learning; the data from the questionnaires administered to the patients at the end of their sessions would reveal differences in terms of patient satisfaction with the two methods of instruction; the data from the behavioral observations would reveal differences between the two methods of instruction in terms of the ease with which the new patients complete their first clinic visit following their class.

8.2 THREATS TO EXPERIMENTAL VALIDITY

Cause and effect are the central concerns in experimental research. The experimenter hopes that the data collected will demonstrate that the

conditions or situations (**treatments**) created or manipulated had some effects on the subjects. Moreover, the experimenter hopes that the results **will generalize,** apply to a population larger than the group of subjects employed during the experiment and ideally apply to large classes of individuals. In the example under consideration here, the experimenters would hope that their results would apply to all new patients who enter the family planning agency involved. Ideally, the results would generalize to all new patients at all family planning clinics. More realistically, however, the experimenters would hope that their results apply to new patients at family planning clinics serving similar populations: similar in terms of demographic and personological characteristics (age, sex, race, native language, IQ, ethnicity, culture, and the like).

Therefore, experimenters must exercise caution that rival explanations cannot logically be offered to explain the meaning of the experimental data. When experimental results are analyzed using inferential statistics (9.3), if statistical differences among the experimental groups appear in the outcome measures (**dependent variables**), did the experimental conditions (**independent variables**) cause those differences, or did some other **intervening or confounding variables** (factors) cause the differences?

Situations or experimental procedures which might lead to rival explanations of the data are called **threats to validity.** When threats to the validity of an experiment muddle the relationship of the data to the research questions, the threats are called **internal.** When threats to the validity prevent the researcher from generalizing the results to populations larger than the experimental subjects, they are called **external.**

8.2.1 THREATS TO INTERNAL VALIDITY

The following analysis of the research design described in the last section is based upon the threats to internal validity identified by Campbell and Stanley (1963).

History. Did some event occur during the course of the experiment which could have affected the results through an effect upon the subjects other than the experimental conditions? Occurrences which pose **historical threats** to internal validity might be things which an entire society experiences, or they might be experienced only by subjects in the experiment. In the case of the hypothetical experiment comparing two methods of instruction, the outcome of the experiment might be affected if the projector

or other audio-visual equipment breaks down frequently, or even a few times, during one or more of the classes. A fire drill, a power outage, or some similar disruption during one or more of the new-patient classes might cause elimination of the results gathered from that particular class at the end of the experiment.

Maturation. This kind of effect includes fatigue, hunger, aging, boredom, and the like. The design under analysis in this section is not likely to suffer from maturation effects because the course of the treatment, the new-patient instruction, lasts such a short time that any maturation effects would probably be equal for all groups. However, if some of the new patients come to their new-patient class, leave the facility, and then return several days or weeks later for their first clinic appointment, their results on the third outcome measure--the behavioral observation questionnaire--might be affected by the intervening time. Consequently one would analyze the results gathered from those subjects who went directly from the new-patient class to the clinic appointment. Or, if many patients at the clinic commonly make two separate visits for their new-patient class and their first clinic visit, then that factor should probably appear as an independent variable in the final analysis.

Testing. If an experimenter gives the subjects a pretest, the subjects may be sensitized to the content of the instruction or other kind of treatment involved in the experiment. Since the experiment in this chapter does not involve a pretest, there is no concern about this threat to internal validity. However, if you do decide to conduct an experiment which involves a pretest, then you might eliminate the testing threat to internal validity by including control groups in the experiment who do not receive the pretest. Consult Campbell and Stanley (1963) for more details.

Experimenters also use **unobtrusive measures** as pretests to eliminate the testing threat. An unobtrusive measure is a test or other data gathering technique which measures the quantity or characteristic of interest but which does not reveal the content of the experimental instruction or renew the subject's knowledge of the material. For instance, suppose that a researcher hopes that a training program will increase new clinic-staff's knowledge of reproductive anatomy, physiology, and contraception. Then, instead of giving a pretest on those topics, one might use a questionnaire which reveals the subject/trainees' prior personal and work experience related to knowledge of those things.

Instrumentation. This threat refers to the validity and the reliability of any tests, questionnaires, or other measurement techniques used during the experiment. The validity and reliability of tests is covered in 6.2.

The behavioral observation technique for assessing the ease with which patients complete their clinic visit is subject to the same kinds of validity and reliability concerns as any other form of measurement. In particular, the reliability of observation techniques depends entirely upon the consistency of the observers or raters who complete the behavioral observation questionnaire. Researchers typically document the reliability of observation techniques by putting their raters or observers through a training period during which agreement about ratings and other kinds of answers recorded during observation is carefully discussed. Next, the raters observe and rate practice sessions or events. Then, the experimenters calculate correlation coefficients (see 9.3.2) for the data recorded by every pair of observers. Thus, if there are three raters, A,B, and C, there would be three correlation coefficients: A versus B, A versus C, and B versus C. Those coefficients are called **interrater reliability.** Those coefficients should exceed .90 and are ideally .95 or higher.

Statistical regression. If you give someone a standardized test, such as an IQ test, and he or she scores very high or very low, psychologists would expect that individual to score closer to the average on the second administration of the same test. This phenomenon is called **regression.** If you give the subjects in an experiment a norm referenced (see 6.1) pretest and posttest, then you might expect some change in their average scores because of regression toward the normed average or mean. Regression is not a concern in the experiment I am examining here, but it becomes particularly significant in experiments when groups are chosen because they exhibit extreme characteristics such as very high IQs.

Selection. In order for experimental results to generalize to larger populations than the ones involved in the experiment, you have to make sure that the manner in which you selected your subjects did not affect the data and consequently the outcome. Also, as indicated in 9.3, independent assignment of subjects to groups for an experiment is a prerequisite to applying inferential statistics to experimental data for comparing groups. Independent assignment occurs if one subject's group assignment does not

affect any other's. If you use random selection (7.3) to choose subjects then you eliminate this threat as long as you independently assign the subjects to groups. In the sample experiment which I am examining here, selection is not a problem as long as patients are randomly assigned to new-patient orientation sessions as they call the clinic and ask for an appointment or appear at the clinic and ask to attend the class. Technically speaking, since the experimental design assigned treatments to already existing groups--the scheduled new-patient orientation sessions--the units of selection and analysis are the groups attending the various sessions. However, in this case, that is of minor concern because the selection procedure does not affect the outcome of the experiment; the two methods of instruction were randomly assigned to the groups, and over the course of a month one could reasonably expect a cross section of typical patients to attend each kind of instruction.

Mortality. When an experimenter loses subjects from experimental groups, experts call this **mortality.** A loss of subjects can threaten the outcome of an experiment if subjects are lost due to characteristics of the subjects which might affect the experimental outcome. For example, if you are doing an experiment to determine how best to reach migrant agricultural workers with family planning services, you will face the constant problem of seeing your subjects move on to other agricultural areas.

Even if subjects are lost for random reasons, reasons which do not affect the outcome of the experiment, mortality can still pose a problem when you begin to analyze your results. Most inferential statistical procedures require as many subjects as possible in order to find important results. Thus if you begin an experiment with small numbers of subjects, mortality can create problems when you want to analyze your data, because your number of subjects drops to a level below that which is acceptable. (See 9.3.)

8.2.2 THREATS TO THE EXTERNAL VALIDITY

When a researcher hopes to generalize the results of an experiment to a population larger than the one from which subjects were chosen, then consideration must be given to making sure that the conditions of the experiment permit those generalizations. When situations or conditions would prevent generalizing to a larger population, experts call them **threats to external validity.** The discussion of these threats below is based upon the work of Bracht and Glass (1968).

Accessible versus target population. The group to which an experimenter hopes to generalize results is called the **target population.** However, only infrequently can an experimenter draw samples of subjects from the entire population of interest. Some smaller group, called the **accessible population,** offers the experimenter a chance to select samples of subjects. If the accessible population is not **representative** of the target population then you cannot easily make the logical arguments necessary to generalize your results from the experimental samples to the accessible population and then the target population. (Representativeness is covered in 7.3.)

In the experiment under examination here, if the results will apply only to the patients at that clinic, then there is no concern about this threat to external validity. However, if the results will be shared with other family planning educators and applied to populations of patients at other clinics, then the accessible population, those patients attending this clinic's sessions, must be similar to the patient populations at other clinics where the results will apply. Since the process compares two methods of instruction, one must make sure that the accessible population is similar to the target population in terms of characteristics which could affect learning: age, educational level, literacy, native language.... Therefore, if I plan to generalize my results, I would need to collect data from my samples regarding these variables so that I could report them to other educators, thus allowing others to determine whether or not the results apply to their patient populations.

Interaction of treatment and personological variables. If an experimenter chooses subjects from an accessible population which has characteristics specifically related to the questions under study, then those characteristics might influence the way the subjects respond to the treatments. Suppose that instead of assigning the typical new patients to the two instructional methods, groups of college freshmen living in dormitories were used. Then, if college freshmen are more literate and have higher IQs than the general patient population, the results might be affected and threatened by those personological variables.

Description of the independent variables. The report on any experiment must include a sufficiently complete description of the treatments to allow others to duplicate or replicate them as exactly as possible. In

order for other educators to know how well the results of my experiment would apply to their patients, they would have to be able to duplicate or replicate the kind of instruction involved. This would probably require making copies of the audio-visual instructional materials and outlines from the lecture or other instructional method available to other educators.

Multiple treatment interference. Sometimes experimenters conduct more than one experiment with the same group of subjects, either simultaneously or at different times. This practice interferes with the results so much as to render them almost completely invalid, particularly when the two experiments are closely related investigations.

The Hawthorne Effect. About fifty years ago, some social scientists visited the Hawthorne (Ill.) Works of the General Electric Company to conduct an experiment, hoping to determine if the lighting which workers had in their work areas would affect production. The workers discovered that they were the subjects of an experiment, and competition among the treatment groups in the experiment developed. Each group was determined to out-produce the others! Obviously, the researchers had difficulty drawing conclusions from their data because their subjects' competitive desire influenced the researchers outcome measure, the number of units produced.

Clearly, if the subjects in an experiment do not know that an experiment is going on, then the Hawthorne Effect is not relevant. Professional ethics dictate that subjects be told that they are engaging in an experiment when there is any reason, however small, to believe that their participation in the experiment might have an adverse effect upon them. In the experimental research examined here, I would have no ethical problem with not telling the subjects they were involved in an experiment to compare two methods of instruction. I would feel comfortable with that because I presume that the two instructional methods are fairly similar and that any differences with respect to knowledge gained or information made available during the new-patient sessions would be discovered and corrected during the patients' visits to the clinic.

Novelty and disruption. If one decides to conduct an experiment using existing groups of subjects and then proceeds to disrupt their normal routine or create novel experiences, those disruptions and novelties can affect the results. This threat to external validity is irrelevant to the experimental example in this section. However, in an experiment comparing

two methods of values clarification for high school students using guest speakers to do the values clarification sessions--rather than the usual classroom teachers--the results may not apply if classroom teachers attempt to use those results later. The guest speakers in the experiment would be novel to the classes, whereas the usual classroom teacher would simply be presenting a new lesson without as much novelty effect.

Posttest sensitization. Sometimes the subjects in an experiment will understand something they were taught during the treatments at the time they read the questions on the posttest. A subject reads a question and has an "Ana!" experience. While the posttest sensitization threat is generally insignificant, in the experiment critiqued here slightly different outcomes might occur without a posttest of new patients. I would expect any such differences to be slight and of minor concern in relation to the other considerations involved in doing the experiment.

Measurement of the dependent variables. As with the earlier concern about being careful to accurately and completely describe the two methods of instruction, sharing the results with others would require making the outcome measures available to other educators. That means that any report written about the outcome of the experiment would have to include an exact description of the knowledge test, the patient satisfaction questionnaire, and the behavioral observation technique. Making these instruments available is important so that others will know exactly what criteria were used in comparing the two methods of instruction, because other educators might have different reasons for selecting one method as opposed to another.

Experimenter bias. Virtually everyone involved in conducting an experiment has prejudices about what they hope the results will show. If an educator conducting the sample experiment on two instructional methods hopes to find that the audio-visual method is as good or better than the lecture method, then s/he may accidentally or intentionally try to affect the results. You can eliminate the experimenter bias by excluding educators who have a vested interest in the outcome from participation in teaching new-patient classes or from making behavioral observations. Since personnel working in the clinic might also have prejudices about the outcome of the experiment, I would want to make sure that I took steps to avoid allowing those prejudices to affect the results. I might do that by asking the clinic personnel to complete the behavorial observation sheet

without letting them know which patients have had which instructional methods. Or, I might solicit the assistance of volunteers, such as nursing students, who will not have any vested interest in the outcome of the experiment because they will not work with its results on a daily basis.

Summary. Essentially, you do not need to be concerned with threats to external validity other than experimenter bias if you do not want your results to apply to situations other than the specific one where the experiment is conducted. The threats to internal validity, however, are important considerations in the design and analysis of any experiment.

8.3 ANALYSES OF THE RESULTS

Section nine of this manual carefully describes the information you will need to make statistical analyses of your experimental results. Statistical analysis is a tool, a method to use in helping to interpret the results, not a substitute for judgments. Any time you interpret the results of an experiment, whether you conducted it yourself or someone else conducted it elsewhere, remember to keep the original experimental problem and questions foremost in your mind as guides to the meaning of the results.

The hypothetical experiment comparing two methods of instruction would have three outcome measures with which to compare the two methods. If an analysis of the results showed that one of the two methods was clearly superior to the other, then the meaning of the results would be clear and their implications for future programming would be straightforward. However, if the two methods showed similar results, then I would need to return to my original purpose for conducting the experiment in order to make judgments about how I would use the results in my planning.

Suppose that the behavioral observation technique and the patient satisfaction questionnaires showed very little difference between the two methods of instruction. Suppose also that the knowledge test given at the end of each new-patient session showed than patients who heard the lecture method scored 5 points higher on the average that those who saw the audio-visual presentation. If the knowledge test had a possible range of scores from 0 to 100, and the lecture group scored an average of 85 while the audio-visual group scored an average of 80, that result may be statistically significant but I would view it to be of little instructional significance.

What would I do if the two instructional methods got similar results on the knowledge test and on the behavioral observations, but patients expressed a clear preference for the lecture method over the audio-visual approach? Should that be the final, determining criterion for making a choice between the two methods? Obviously, the answer depends upon what I would have to exchange in terms of resource management, educator burn-out, and staff morale if I make my choice in terms of the patients' preferences. The value systems of the people who interpret the experimental results would go further toward determining the final decision than the results themselves.

I would have a strong tendency to present my experimental results to other professionals or to members of the board of directors, to people who do not have an interest in the final decision, and get their advice about how to proceed. In some family planning agencies a high level administrator or a committee of the board of directors might reserve the right to make final decisions about the meaning of the experimental results.

A post script. Experimental research throughout the social sciences has been subjected to untold amounts of criticism. There seems to be enough clear evidence to justify the assertion that most experimental research conducted by behavioral scientists has limited practical value, and that it may even be worthless. Hopefully, the discussion of threats to validity contained in this section and the basic information about data analyses in the next section will give you some valuable tools and information to use when you read and critique other people's experimental results.

SELF-TEST ON SECTION EIGHT

You will probably gain the maximum benefit from this self-test if you write your answers on a separate sheet of paper and then compare them with the answers on the pages following the self-test.

1. State three criteria for deciding whether or not to conduct an experiment in an educational setting in terms of whether a given situation justifies the expenditure of time and other resources.

2. Does every experiment have to have a control group or groups? Why or why not?

3. In each of the following situations, identify the threat to the validity of the experiment in question and state why it is a significant threat.

   A. During one of the educational sessions designed for the experiment, a fire drill is called and the participants must vacate the building.

   B. A knowledge test with test-retest reliability of .73 (see 6.2.2) fails to show a significant result between two experimental groups.

   C. An experiment with two experimental treatment groups is conducted over the course of one school year for seventh graders but no control group is included in the experimental design.

   D. The results of an experiment involving a posttest do not seem to hold true when the researcher attempts to apply the experimental techniques to groups who do not take the posttest.

   E. An experimenter begins an experiment with 20 subjects in each of two treatment groups but loses 8 subjects from one group and 10 from the other during the course of the experiment.

   F. An experimenter hopes to generalize results to all females aged 13 to 19 and has access to subjects for the experiment who range in age from 15 to 17.

   G. An experimenter wants to do research on a teenage population and learns that another researcher is conducting an experiment with exactly the right kinds of groups of subjects. The two agree to use the same groups of teenagers for their two experiments.

   H. In an experiment to compare two methods of teaching contraception to high school sophomores, the students learn that they are subjects in an experiment and place bets on which group will score highest on the final examination in the course, the outcome measure for the experiment.

I. You read about the results of another family planning educator's experiment and want to apply those results to your setting, but the research report does not specify what kind of test was used as the outcome measure (dependent variable) for the experiment.

J. Two educators decide to settle a dispute about two different methods of giving high school talks by conducting an experiment to compare the two techniques.

4. You have just conducted an experiment to compare two different methods of teaching new-patient classes. You used a test of knowledge on the material covered in the new-patient education sessions as your outcome measure. When you calculate the average scores for the two different groups, their mean scores differ by 3.5 points on a test with a range of possible scores from 0 to 85. You use an inferential statistical technique (see 9.3) to compare the average scores and the result is statistically at the level of only 1 chance in 50 of getting the result you observed by random chance. Do you conclude that the method which had the higher mean is clearly the superior method of instruction and proceed to use that method? Why or why not?

5. At what point in an experimental research project do you begin to consider which statistics you will use to analyze your results? Why?

ANSWERS TO SELF-TEST ON SECTION EIGHT

1. (a) Does the question involved lend itself well to an experiment or would a survey approach do just as well? (b) Has anyone else conducted a similar investigation, or one which could answer the questions in part, based upon a search of the literature on family planning and sexuality education? (c) Is the question of sufficient interest and import? (d) Do you have access to settings and subjects which would enable you to conduct a valid experiment? (e) Will the results be useful to any other professionals in family planning education?

2. No. Sometimes a control group is the only method for handling threats to the validity of experiments, but sometimes the threats which necessitate a control group do not apply to a situation. In other instances, advanced statistical techniques enable you to avoid using a control group.

3. A. This historical effect might influence the final results because of the disruption experienced by one group but not by the other group(s) involved in the experiment.

   B. Instrumentation; this threat is extremely significant because you cannot know whether your treatments really made a difference or whether your test was so poor that it could not detect true differences.

   C. Maturation; in addition to the effects caused by the two different methods of instruction for the seventh graders, the subjects would experience changes which might affect the results because they mature a lot during their seventh grade year in school.

   D. Posttest sensitization; the subjects of the experiment understood or learned the material while studying for or taking the posttest, but other people who experience the same treatments later do not have the same learning experience of the posttest. Thus, the people who get the same treatments later fail to learn the same things.

   E. Mortality; this could affect the results if those who leave drop out for reasons related to the experimental conditions. This might also create problems because of small numbers of subjects when the results are analyzed by statistical procedures.

   F. Accessible population does not match the target population and therefore the results may not generalize to the target population.

   G. Multiple treatment interference; neither researcher could tell how the presence of the other's experiment affected their results.

   H. The Hawthorne Effect; the subjects' scores on the final exam will reflect not only the differences between the two teaching methods but also their motivations to win the bet, making the results inapplicable to other situations.

I. Measurement of the dependent variables; you do not know if the outcome measure is aimed at the same things you are interested in for your situation.

J. Experimenter bias, a chronic problem in any research. In this case you will not know if the experimental results fairly reflect the differences, if any, between the two methods, or whether the results reflect the competition between the educators, the educators' submission to temptations to "fudge" data, or just exactly what will happen from a validity standpoint.

4. No. While the result is statistically significant, it is not of sufficient importance to be a determining factor in choosing one method over another. Your original experimental research question(s) should offer a better set of guidelines, as well as the motivation behind the comparisons: cost, morale of staff, quality of instruction, ...

5. As early as possible. Not every set of experimental data has a matching statistical analysis procedure, and the manner in which subjects are chosen, groups assigned, and the number of subjects decided upon will often limit your choices for data analysis. (See **section nine.**)

section nine:
ANALYZING DATA

After reading this section and learning the material in it you will be able to:

1. Define the following sets of terms and state the differences among them:

   a. inferential statistics
   b. descriptive statistics

   1. parametric statistics
   2. nonparametric statistics

   i. levels of measurement
   ii. nominal
   iii. ordinal
   iv. interval
   v. ratio

2. Create frequency distributions, grouped frequency distributions, tables, charts, and graphs.

3. Calculate percentages and percentages of change (increase or decrease), and calculate the amount of a given quantity which corresponds to a specified percentage of that quantity.

4. Define, calculate, and interpret means, medians, and modes.

5. Define and describe the importance of skewness.

6. Define and calculate ranges, variances, and standard deviations and interpret the meaning of a standard deviation.

7. Define the meaning of a correlation coefficient and calculate and interpret Pearson's r and Spearman's r_s.

8. Discuss various statistical inference problems in a language which will enable you to communicate with a consultant who is familiar with statistics and statistical tests.

If you are familiar with the material in this section, you may want to skip to **section ten.** If you are not sure, or if it has been a while, you may want to use the self-test at the end of this section as a review or as a check to see which portions of this section you need to review.

NOTA BENE: This section is an **extremely basic** discussion of elementary statistical terminology and procedures. If you have no background in statistics, or relatively little, then you would be wise to consider enlisting the advice of a consultant who knows a good deal about statistics. If you have studied some of these topics but have not worked with them in a while, you may need the help of a basic statistics text and a few hours of review. The **Glossary** may prove helpful as you read this section.

section nine: ANALYZING DATA

"Most people use statistics the way a drunk uses a lamp post--more for support than for illumination."

--Anonymous

Data analysis, commonly called statistics, has a reputation for a propensity toward Greek--a series of Greek letters strewn across a page followed by explanations in an English which may as well be Greek. This section will avoid those pitfalls. Any equation for calculating the statistics will be accompanied by instructions which will reduce the process to steps consisting of addition, multiplication, subtraction, division, and the calculation of square roots (explained later). Any of these arithmetic operations can easily be performed with a ten-dollar hand-held calculator, although an adding machine or calculator with printed tape of the numbers entered and the results might help verify lengthy addition.

In keeping with the philosophy adopted throughout this manual, this assumes little prior knowledge on your part. I intend to convey a basic guide for calculating those numbers which might frequently prove valuable, with some discussion of the more advanced tools so that you can communicate with consultants and peers about possible options beyond the scope of this section but nonetheless potentially useful.

Statistics are just that--tools. They are simply numbers calculated from the theories of mathematics and based upon the laws of probability. As with any tool--a drug, a hammer, a laser, a scalpel, or a bulldozer--they have their appropriate uses and their abuses. They are neither intrinsically good nor bad; their reputation should fall or fly upon the reputation of their users and uses.

The calculations which yield statistics use whatever numbers the calculator feeds into the process, whether appropriately or inappropriately, whether accurately or erroneously. The formulae, procedures, and numbers have no memory; their outcome depends upon their use. Thus, this discussion begins with a description of the **levels of measurement**, the degree of fineness with which the raw data are gathered as numbers. Then it will turn to the various **descriptive statistics**, numbers calculated to summarize the characteristics of a body of raw data. They include

measures of central tendancy, such as **means** (averages), **medians**, and **modes**, and measures of dispersion, such as **range**, **variance**, and **standard deviation**. Finally, various kinds of **inferential statistics**, numbers calculated to quantify conclusions about data, will be presented. There will also be a section on the various ways of presenting raw data.

There are, of course, numerous texts on statistics. Huck, Cormier, and Bounds (1974) may prove most helpful for people who hate mathematics but who need an understanding of reading and interpreting research and evaluation reports and the statistics in them. Of the several-score basic statistics texts, Runyon and Haber (1967) took pains to present the material with as few assumptions as possible about the reader's knowledge of algebra, calculus, and matrix alegebra--to the point that a person who does not know those subjects can still read the text. For those who want a complete intermediate text for reference, explore Glass and Stanley's (1970) text. Miller (1977) includes various calculation strategies and an extensive bibliography. Siegel (1956) wrote such a thorough and readable text on nonparametric procedures (explained in 9.3.1) that it has attained classical status as the unrivaled source for beginners. If you would like more information about the theory presented here, or more about the kinds of tests described, you can refer to any of these texts. The statistical booklet by Morris and Fitz-Gibbon (1978), from **The Program Evaluation Kit** by the same authors, might also prove helpful.

9.1 LEVELS OF MEASUREMENT

Scientists in disciplines from physics to anthropology have known for centuries that the numbers they call data vary with respect to how finely they measure quantities. But not until 1946, when Stevens published a treatise on this question in the journal **Science**, did anyone attempt to categorize the various levels at which data are measured: **nominal**, **ordinal**, **interval**, and **ratio**. The understanding of these concepts will prove inescapable later in this section, although you may wonder why they are of any interest other than academic. In fact, an understanding of the levels of measurement forms the basis for the majority of initial decisions about the kinds of statistics you will use to analyze data.

Nominal. Data measured at this level place objects, people, or other entities into categories. **Data measured at the nominal level permit simple statements such as that A is different from B.** However, nominal

variables (quantities being measured in different amounts or qualities) permit only statements to the effect that A is different from B, but not that A is greater than (symbolized here as >) B nor that A is less than (<) B. Sex, race, religion, origin, socioeconomic status, class, location, group membership, and all other categorical measures are nominal.

Ordinal. This level permits descriptions in terms of rank. **An ordinal-level variable permits you not only to make statements that A is different from B, but also that A is either greater than or less than B.** But the ordinal level does not imply that we can discern how **much** greater or less one category is than another. All forms of ranking without any indication of how far apart the individually ranked entities lie are ordinal. The current standings in the American League's Eastern Division of Major League Baseball are ordinal. If we examine the standings without won/lost records or winning percentages, then we know that one team leads the others, another is in second place, and so forth, but not by how much. The results of a respondent's work at a Q-sort (see 5.2) are ordinal data. Rank in class, final grades in a course given as A, B, C, D, F, and ranks in the military are ordinal.

Almost all questionnaire data are ordinal, unless they are ages or one of the variables discussed below as interval or ratio. Response categories on Likert-type items, rating scales, and some check lists' results are ordinal. Semantic differentials, however, yield nominal data.

Interval. At the interval level, we can discern **not only the direction of differences but also their quantity.** However, the units of measurement on an interval scale have an arbitrary zero point, as with Fahrenheit and Centigrade temperature scales. The interval of degrees between 30°F and 50°F is the same as between 200°F and 220°F, but 0°F was arbitrarily established and does not represent a true absence of heat. Percentages of effectiveness for contraceptive methods achieve at least interval-level measurement. Scores on achievement and IQ tests are usually treated as interval data, although some purists and mathematicians would deny the validity of that practice. Scholastic Aptitude Test (SAT) scores are mathematically ordinal because a difference of 100 points is not the same between 450 and 550 as it is between 550 and 650. College admissions personnel often treat them as interval data. (SAT scores are not interval data because they are derived by translating a raw score into a score based

upon the mean and standard deviation of the data, a form of z-score, or standard score.) Interval-level data allow direct transformations from one interval scale to another interval scale for the same quantity, such as converting Fahrenheit temperature to Centigrade. However, an SAT Score of 450 represents the 41st centile; 550, the 69th centile; 650 is at the 93rd centile; while 750 is at the 98th; clearly intervals of different size.[1]

Ratio. Data at this level are measured on a **scale with a zero point truly representing an absence of the measured quantity.** The Absolute or Kelvin temperature scale is ratio because, as far as physicists can determine, 0°K represents the complete absence of heat. Height and other measurements of length, time (age, duration of an event), weight, and the balance in your checkbook are ratio-level scales. The ratio level of measurement was so named because the absolute zero point permits calculation of ratios of the quantity. For instance, 40°F is not twice as hot as 20°F, but 200°K is twice as hot as 100°K, and $20 is twice as much as $10. Ratio-level measurement is rarely if ever used in the social sciences, with the exception of quantities like time, distance, and similar measurements borrowed from the physical sciences.

What the differences mean. These differences in levels of measurement are significant because one can perform mathematical calculations which have some meaning with interval and ratio level data, but there are less of those options for calculating statistics with ordinal and nominal data. For example, you might calculate the average of a set of test scores, but you would never calculate an average zip code. Zip codes are nominal data. Although they take the form of five-digit numbers which you could enter into a calculator and average, it would be nonsense. Nominal and ordinal data are sometimes manipulated by additions, subtractions, multiplications, and divisions, but the meaning of the statistics thus calculated is based upon the assumption that pure arithmetic operations cannot be performed on the data. Mathematical laws are taken into account in the methods used to interpret the results of the calculations.

---

[1]A **centile** (also called a percentile) is a way of translating test scores into numbers to allow comparisons among the test takers. A centile is the percentage of people who took a test and scored below that test score. A score at the 50th centile is that score representing the point where 50% of the candidates scored below that score (and the other 50% above). A score at the 84th centile is above 84% of the candidates and below the other 16%.

Nominal data permit little more than counting numbers of persons or things in each category, although it is possible to draw conclusions by manipulating the counts. Ordinal data may yield much more information than nominal, principally through **nonparametric** statistics. At the two higher levels of measurement, all common mathematical operations, such as the operations one performs when measuring areas, speed, accelerations, and the like, are permissible.

9.2 DESCRIPTIVE STATISTICS

This section contains the common methods for presenting raw data in reports and for using the statistics which one can calculate with all four levels of data.

9.2.1 PRESENTATION OF RAW DATA

When you have some raw data--test scores, counts of events, number in an audience, questionnaire responses--there are various methods available for presenting the data.

Throughout this section and the next, I will need sample data to illustrate the analytical procedures which I will present. Therefore, the **Sample Data Bank** contains a set of fictitious and meaningless numbers to which I will refer from time to time to illustrate concepts and statistical calculations.

Tables. Social scientists present data in tables not only because they consume less space than narrative presentations, but also because they allow the reader to inspect the data more quickly and to still do so thoroughly. Tables of data should be as brief as possible and offer an attractive presentation.

Sample Data Bank

	Columns	
Rows	A	B
1	22	19
2	24	32
3	24	26
4	26	17
5	30	31
6	31	36
7	31	34
8	31	50
9	39	41
10	42	60
11	46	45
12	52	50
13	52	39
14	55	61
15	55	62
16	55	38
17	55	55
18	59	70
19	61	66
20	70	58

The data in this table are fictitious and will serve only as a repository of raw data for calculating examples of statistical quantities in the remainder of section nine.

Avoid ribbon-and-pencil tables--long strings of numbers in one row or column. Combine several kinds of information about the data in one table whenever possible. The table should be labeled succinctly but informatively and numbered or lettered if there is more than one, to facilitate reference to them in the text of a report.

The data such as those contained in column A of the Sample Data Bank lend themselves to a tabular presentation as a **frequency distribution**, or a report of data by categories with the number of times each category appears in the raw data. If the data in column A were scores on a final exam in a course for education volunteers, you might present them as the following frequency distribution:

SAMPLE TABLE A: Distribution of Final Exam Scores

Score	Frequency
22	1
24	2
26	1
30	1
31	3
39	1
42	1
46	1
52	2
55	4
59	1
61	1
70	1
Mean = 43.0	n = 20
Median = 44	

Each number in the column of Sample Table A labeled frequency is the number of trainees who earned that score. The symbol n at the bottom of that column is the standard representation for the total number of data points, as when reporting the total number of questionnaires completed, total number of attendees at a session, etc. The mean and median appear at the bottom of the table as well, but the mode is not reported because it is obvious to any reader upon inspection of the table. In this case the presentation of the raw data as a frequency distribution saves relatively little space; when there is a very large number of data points a frequency distribution saves a great deal of typing and space. (See mean, median, and mode--9.2.3.)

The body of the table might have looked like this:

Score range	Frequency
22-30	4
31-40	5
41-50	2
51-60	7
61-70	2

which is called a **group frequency distribution.** The space-saving advantages are obvious, but loss of specificity occurs as well. This kind of presentation is most appropriate when the categories for the groupings have some special significance, as when you want to distinguish between poor, average, and superlative performances.

One can calculate or verify another's calculations of descriptive statistics quite readily from Sample Table A, but the same is not true of the grouped frequency distribution.

Graphs. Some data, particularly the counts within categories of a nominal variable, lend themselves well to graphic presentation. And when a visual image will help convey a presentation more dramtically, a graph does the job much better than a table of numbers.

Histograms, the ever popular **bar graphs,** are easy to construct, but there are a few common ethics associated with such presentations. Picture a bar graph arranged so that the bars run up and down the page. The first rule is that the tallest bar should not exceed 80 to 90 percent of the total height of the graph, to avoid giving a false impression that the largest category has run off of the page, unless that is truly the case (e.g., if there is a sudden and incredibly remarkable increase in patient load, say a 300 percent increase). The units of height should be uniform throughout the graph; do not make the distance between 0 and 10 half the size of the distance between 10 and 20. To accommodate situations where you want to show the difference between nearly equal but large quantities, use broken bars, bars with a jagged space near their base. Bars may be contiguous or separated by white space, the latter offering the clearer presentation.

Frequency polygons are fancily-named line graphs. They are like bar graphs, only they have points, connected by straight lines, located at the top of where the bars would go. Do not use a curved line to connect the

points. Frequency polygons offer a more dramatic presentation of peaks and valleys, particularly when there is a clear and consistent trend across the graph.

Bar and line graphs can be combined to save space, as with tables. Care should be taken to use contrasting lines or bars for different quantities, with attention to the loss of color and definition on photocopy machines.

9.2.2 PERCENTAGES

The word "percentage" comes from the Latin **per centum,** or "per one hundred." A percentage is a fraction expressed as a decimal and multiplied by 100. Think of percentages as expressions of parts per one hundred of a total. To calculate what percentage of B a quantity A is, divide B into A, write the results as a decimal, then multiply by 100.

Example: What percentage of our 9,000 pieces of literature did we obtain free of charge if we got 2,750 free pieces of literature.

(Oh no! Word problems!)

Answer:

$$\frac{2,750}{9,000} = .3055;\ .3055 \times 100 = 30.55\%$$

To convert from a percentage back to a whole number, divide the percentage by 100 and multiply the total quantity by the resulting decimal.

Example: If 22% of 8,534 patients last year were teenagers, then how many teenage patients were there?

Answer:

$$\frac{22\%}{100} = .22;\ 8,534 \times .22 = 1,877.48$$

Obviously, there was not a .48 patient there; you would round to 1,877. The rules for rounding specify that fractions less that one half are rounded down, more than one half rounded up, exactly 1/2 is rounded up in odd numbers, down in even ones. Thus 77.5 becomes 78 but 78.5 also becomes 78.

Can a percentage be more than 100%? Obviously you cannot have more than 100% of a total quantity; you can divide a pie all day and still only have one pie. But sometimes one quantity is larger than another and the larger one is compared to the smaller as a percentage in excess of 100%. For example, suppose your friend earns $18,000 a year and you earn $15,000. Your friend's salary is:

$$\frac{\$18,000}{\$15,000} = 1.2;\ 1.2 \times 100 = 120\%$$

or your friend's salary is 120% of yours.

Percentage change. Planners and evaluators frequently need to calculate percentage change, also called **percent increase or decrease,** depending upon the direction of the change. To calculate a percentage change, subtract the new quantity of something from the old and divide the result by the old (and multiply by 100 to get percentage).

Example: This year we increased our numbers of speaking presentations to 511 from the 476 we did last year. What was the percent change?

Answer:

$$\frac{511 - 476}{511} = \frac{35}{511} = +.0685;\ \times 100 = +6.85\%$$

The plus sign before the percentage indicates a percentage increase.

In some cases you may want to know how much more of a quantity you would have if you increased or decreased it by a certain percentage. To do that, multiply the current quantity by the percentage (expressed as a decimal) and add the result to the current quantity.

Example: (Here's one you'll like!) Suppose your boss announces that you will receive an 8% raise effective on the first of the month. Your currently salary is $12,000 a year. How much of a raise in salary per year are you going to receive?

Answer:

$$\$12,000 \times 8\% = \$12,000 \times .08 = \$960$$

If you want to directly calculate the new quantity itself, multiply by 1 plus the percentage expressed as a decimal. In the last example, your new salary would by $12,000 X (1 + 8%) = $12,000 X 1.08 = $12,960.

9.2.3 MEASURES OF CENTRAL TENDENCY

Measures of central tendency are statistics which reveal information about the location of the center of an array of data. Herein I will discuss the three most common ones, **mean, median,** and **mode,** and make a comment about a concept called **skewness.**

At this point I will begin a practice which I will follow with all of the statistics presented in the remainder of this section. I will define the statistic and give a formula for it, if there is one, for the benefit of those who can make use of the formulae. Then I will discuss appropriate

and inappropriate uses of the statistic, followed by a step-by-step procedure for calculating it, reduced to operations such as multiplication and addition. Finally, I will give guidance as to its interpretation with examples.

Mean. This statistic is also called the **arithmetic mean** (to differentiate it from a monster called **harmonic mean**) and is nothing more than an average. Statisticians typically use X to symbolize one of the several data points to be averaged and $\overline{X}$ (read X - bar) to represent their mean:

$$\overline{X} = \frac{1}{n} \cdot \sum_{i=1}^{n} X_i$$

in which n, as before, is the number of quantities being averaged, the dot ($\cdot$) means multiply, X_i represents the data points, and Σ is the Greek **sigma**, and has been used for centuries to indicate summation.

Uses: Means may be calculated on any interval or ratio data. Some social scientists calculate them for continuous data, that is, data which can theoretically take on any value, rather than just a small number of discrete categories. Means are never calculated for nominal data; such a quantity would be meaningless.

Calculation: To find a mean, add the numbers you want to average and divide the sum by the n of data points. The result is a mean.

For example, the sum of the data in column A of the Sample Data Bank is 860 since there are 20 numbers, $\overline{X} = 860 \div 20 = 43.0$.

If you need to calculate a mean from a frequency distribution (see 9.2.1), then you can rewrite the frequency distribution to make a complete list of the scores for addition. Or you can simply multiply each of the scores by its frequency and add those figures to arrive at the same total, divide by n and get the mean.

Interpretation: A mean is not necessarily in the "middle" of a distribution of scores; it is in the middle in the sense that if you add all of the differences from the mean, accounting for positive and negative distances as being above and below the mean, you will get zero. (See **skewness** below.)

The mean is the most useful measure of central tendency because it can be used to calculate inferential statistics (9.3) for comparing two groups or the same group on two variables (measures). The mean, if correctly used, has the properties necessary to allow application of parametric statistics (see 9.3.1).

The mean and the median determine the direction of skewness, and the amount of skewness can lead to important interpretive results for evaluative purposes.

Median. The score which represents the middle of a distribution, above and below which equal halves of the scores fall, is the median. You find the median by inspecting the distribution. Writing the scores as a frequency distribution or ranking them from highest to lowest will ease that process.

While there are various formulae for calculating medians, they generally consume more energy than they save. If you have a small set of scores, such as 10, 13, 14, 18, 22, then the score in the middle of the group is the median; in the above distribution, 14. Note that the mean of the five scores above differs from the median; the mean = 15.4, a reflection of the fact that the two higher scores differ more from the median than the two lower scores.

Now, examine the data in column A of the Sample Data Bank. Since there are 20 scores, the median falls halfway between the 10th and 11th scores. Since the data in column A are listed in order, the median falls between the scores in rows 10 and 11, or between 42 and 46. But what lies halfway between 42 and 46? Sum them and divide by 2;

$$\frac{42 + 46}{2} = \frac{88}{2} = 44, \text{ the median.}$$

In some cases, things will not be quite so easy. What happens if the data in column A were not so straightforward, say if there were scores of 42 in both the 9th and 10th ranks with a 46 in the 11th spot? Follow the same procedure, add the three numbers, and divide by 3 this time, yielding a median of $\frac{42 + 42 + 46}{3} = 43.33$

In general, follow the rule of summing the scores above and below the median point for distributions with even numbers of scores, and then divide by the n of scores in those spots; i.e., average the scores clustered immediately above and below the median point.

Uses: Medians may be calculated for ordinal, interval, or ratio data, although they make very little sense for purely ordinal data. For instance, the median rank out of nine ranks, with no ties, will always lie at the fifth rank. However, if you have 50 soldiers varying in rank from

private first class to sergeant major, discussing their median rank might reveal something of their relative distribution among the ordinal data of US Army ranks.

Mode. The mode is simply the most frequently occurring score in any distribution. You indentify the mode by inspecting the distribution, and you can state a mode for any kind of data. In the case of small distributions, say 10 scores or less, the mode reveals very little. Consider the scores in column A of the Sample Data Bank. Since there are 4 scores of 55, and that is the most frequent, that is the modal score. The mode for column B is 50 because that score appears twice, and no other score appears as frequently. If another score in column B had a frequency of 2, then column B would be called **bimodal.**

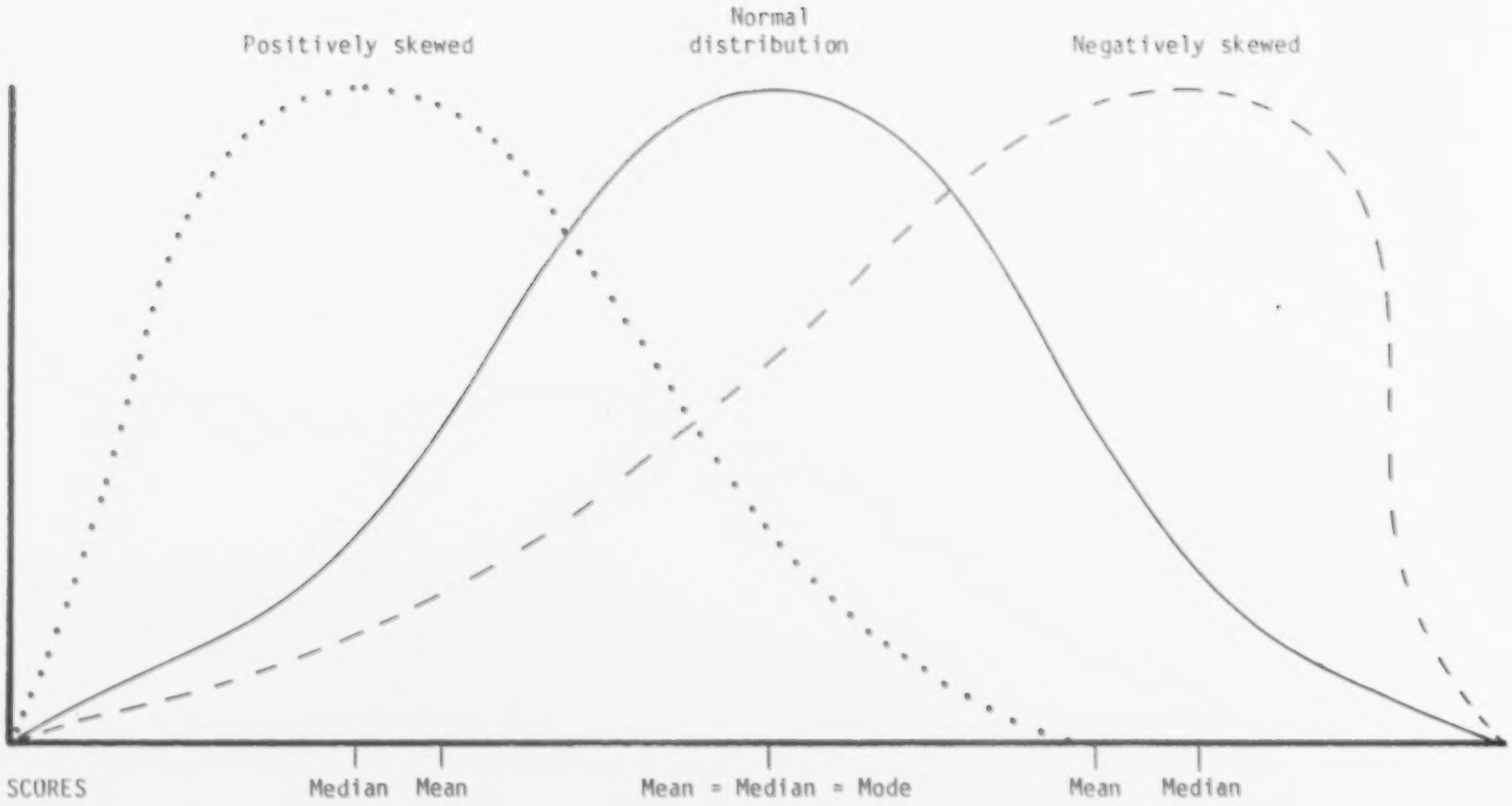
Reporting the modal frequency in a distribution reveals the most when the mode differs greatly from the mean or median. For example, in Zelnik and Kanter's (1972) famous study of teenage sexual behavior, they found that about half of their respondents had had sexual intercourse at least once, but that the modal frequency of sexual intercourse in the month prior to their surveys was zero times. Thus, the data revealed that although the teenagers were sexually experienced, they were not engaging in sexual relations with much frequency.

A mode is the only legitimately reportable measure of central tendency for a nominal variable.

Skewness. Consider the curves in **Figure 1: Normal and Skewed Distributions.** The familiar bell-shaped curve in the figure represents what scientists call the normal distribution, because the curve is symmetrical about the center and has identical mean, median, and mode. Theoretically, various characteristics of human beings and other observed phenomena in nature exhibit normal distributions: size (height, weight), intelligence, strength, mechanical aptitude, artistic aptitude, longevity, and the like.

However, some distributions are skewed--the curve leans to one side or the other--and the mean and median are not identical. Skewed distributions occur in nature, such as the distribution of skin tone for human beings, which is skewed toward the darker tones. Sometimes a distribution exhibits skewness when you examine a subset of the entire population, such as the

FIGURE 1. NORMAL AND SKEWED DISTRIBUTIONS



Notice that the normal distribution, represented by the solid line, has identical mean, median, and mode and that it is symmetrical about a vertical line drawn through the mean. The other two distributions are not normal because they do not have the characteristic bell-shape and because their means and medians are not the same.

184

distribution of intelligence quotients of college students. You would expect to see very few college students with IQs in the lower end of the spectrum, and a larger-than-average number of persons with above average to genius level IQs, yielding a negatively skewed distribution. (The labels "positive" and "negative" are arbitrary and refer to the direction of the distribution's "tail" along the standard number line; positive toward the right, negative toward the left.)

If distribution of scores exhibits marked skewness, then you know that your population or group is heterogeneous. Thus, you will need to make some decisions regarding the level at which you will aim your materials. For example, in writing this manual, I address an audience which includes many experienced and sophisticated professionals, but which is composed mainly of relative beginners. I know these characteristics of my audience from a needs assessment I conducted and from talking with numerous family planning educators. Consequently, I had to decide whether I could appropriately make assumptions about what you, as the reader, know, and decided that the experienced educators could more easily stand a review of some material than beginners could seek references to fill gaps left by any assumptions I might make. Thus, I may bore some of the readers represented by the tail of the skewed distribution, but I attempt to avoid making assumptions about prior knowledge whenever possible.

9.2.4 MEASURES OF DISPERSION

From time to time, you may want to know not only where the center of a distribution lies, but also how widely dispersed the scores are in the distribution. Descriptive statistics which quantify the extent of the spread in a distribution are called **measures of dispersion: range, variance,** and **standard deviation.**

Range. The range of a distribution refers simply to the two most extreme scores, which you can identify by inspecting the distribution. In the Sample Data Bank, column A, the range of scores is 22 to 70.

You can report a range for data measured at any level from nominal to ratio. For nominal and ordinal variables, the range is the **only acceptable measure of dispersion to report, because the others require the use of arithmetic operations on the data,** operations inappropriate for ordinal or nominal data. Once again, I want to acknowledge that social scientists

often treat ordinal data which is "almost interval" as if it were interval and calculate other kinds of measures than ranges. However, it makes no sense to calculate variances of standard deviations for some kinds of ordinal data (such as ranks in the American League East) and no sense at all for any kind of nominal data.

Variance. Variance and its square root[1], standard deviation, are the two most commonly used and reported measures of central tendency. Unfortunately, they often appear inappropriately and abusively in social scientific and educational practices.

Statisticians typically use the symbol σ^2 (read "sigma squared") to represent the variance of a population and s^2 (s squared) for the variance of a sample drawn from a population. Here, use of s^2 will mean variance because it requires less changing of elements in the typewriter.

The formula:

$$s^2 = \sum_{i=1}^{n} X_i^2 - \frac{1}{n} \cdot \left(\sum_{i=1}^{n} X_i\right)^2$$

Uses. The variance reveals the spread of scores in a distribution. A large variance indicates that scores in the distribution vary widely from the mean, while a small variance indicates that the scores cluster nearer to the mean. As will become clearer in the next few paragraphs, the standard deviation is more often used for descriptive purposes than the variance.

You may appropriately calculate a variance on interval or ratio data, with the usual caveats about calculating variances for ordinal data. As before, nominal data cannot yield meaningful variances; it makes no more sense to calculate the variance of a list of zip codes than it does to calculate their mean.

---

[2]The square root of a is b if $b^2 = a$; that is $b \times b = a$. The symbol for the square root of a number, say x, is $\sqrt{x}$, or $x^{1/2}$ as an exponent. For example, $\sqrt{4} = 2$ because $2 \times 2 = 4$. Square roots are not always whole numbers; $\sqrt{5} = 2.306...$, a fraction which goes on infinitely without repeating itself, or any other pattern of its digits.

Interpreting a variance requires experience and judgment, or the application of inferential statistical procedures (9.3). I will discuss arithmetic operations on the data, operations inappropriate for ordinal and interval data. (For nominal data one reports the range of frequencies observed in the categories. For instance, suppose you have counted the numbers of new patients who gave each of the various zip codes in your community with their addresses. Then, you might report that "the most frequent zip code was 10036 with 142; and the least, 10019 with only 6.")

Once again, I want to acknowledge that social scientists often treat ordinal data whch is "almost interval" as if it were interval and calculate other kinds of measures than ranges. However, it makes no sense to calculate variances or standard deviations for many kinds of ordinal data (e.g. the ordinal standings in the American League East) and no sense at all for any kind of ordinal data.

Interpreting variances is discussed under interpretations of standard deviations, below. Standard deviations are used to describe and understand dispersions more often than are variances.

Calculation. To calculate a variance of a group of scores, represented by X's in the following description, make a table like this one:

A. Scores (X)	B. X^2 (or, X times X; score multiplied by itself)
_____	_____
_____	_____
_____	_____
_____	_____
_____	_____
_____	_____
Total _____	Total _____

Then take the following steps:

1. Add up the numbers in both columns.
2. Take the sum in column A and multiply it by itself, square it.
3. Take the number you got in step 2 and divide it by n, the number of scores you have.
4. Finally, subtract the result you got in step 3 from the sum of column B; the result is your variance.

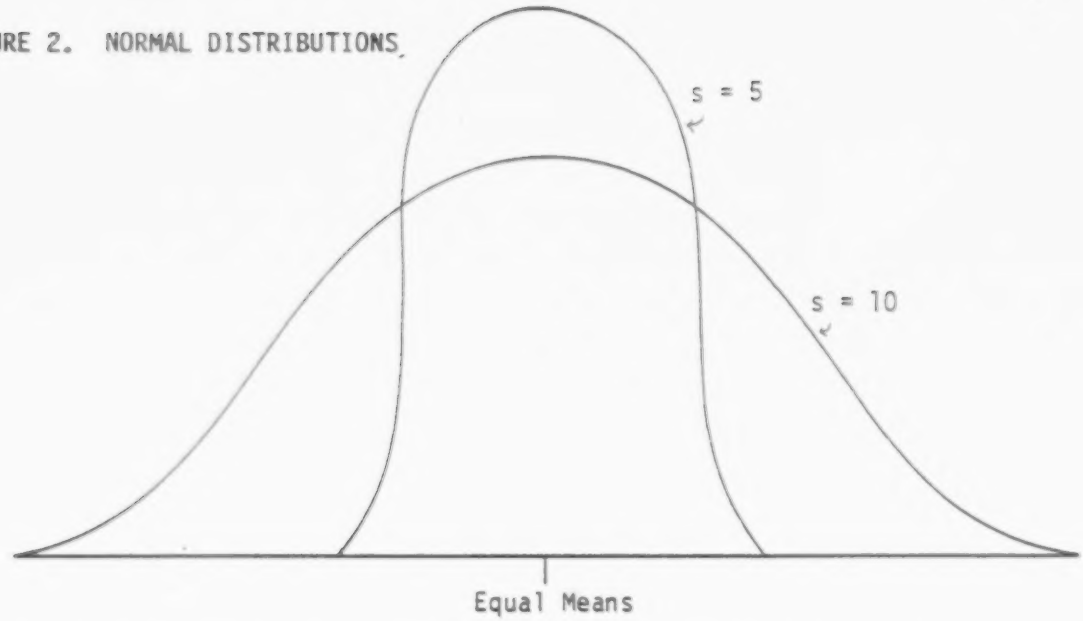
If you would like to practice, use the data in column A of the Sample Data Bank. Your calculations should yield a variance of 206.30.

Standard deviation. If you calculate a variance and then calculate the square root of the variance, you have the **standard deviation**, or s. (Calculating a square root with paper and pencil is an abominable task, and hand-held calculators nearly always have a square root key.)

Figure 2: Normal Distributions, illustrates the significance of standard deviations in terms of normally distributed sets of scores. Part A of Figure 2 shows the difference between two normal distributions with identical means but different standard deviations. The figure clearly shows how much narrower a normal distribution becomes when the standard deviation is relatively small.

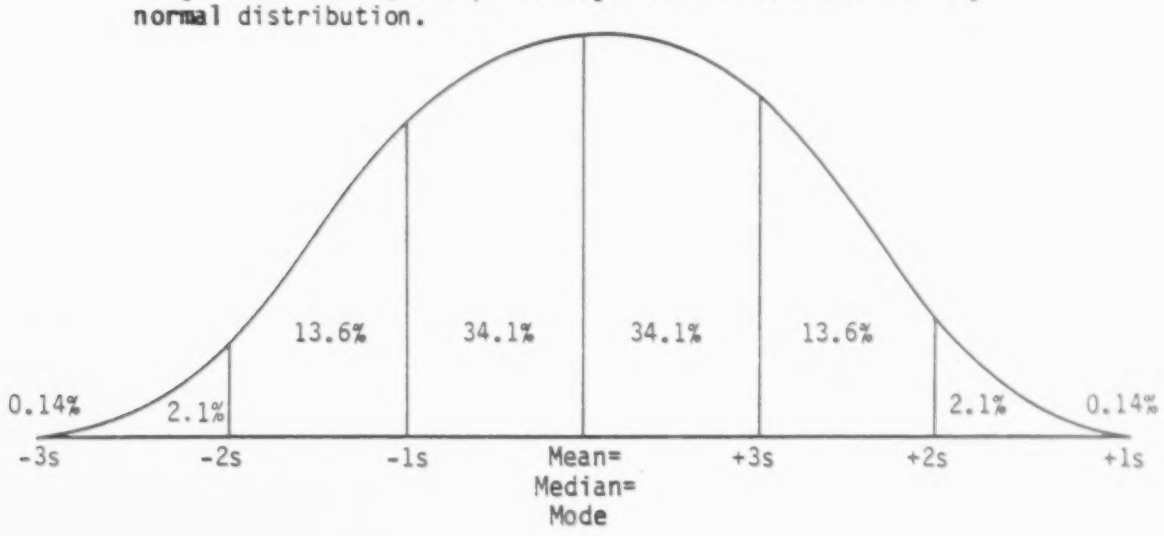
Part B of Figure 2 shows why standard deviations receive so much attention from psychometrists and statisticians. For any normal distribution, as the figure shows, 34.1% of the scores in the distribution fall between the mean and +1s, or one standard deviation above the mean. The same holds true for the band of scores between the mean and -1s; 34.1% of the scores fall in that band. These statements hold true for all normal distributions, but since not all distributions are normal, **they do NOT hold true for all distributions.**

FIGURE 2. NORMAL DISTRIBUTIONS



A. This figure shows the difference between two distributions, both normal with equal means, but with different standard deviations. Notice how the distribution with the smaller **s** is much taller and more clustered around the mean.

B. The percentages in the figure below indicate what percentage of the scores fall within each band of the distribution. While the scores at the ends of the bands will vary with the value of **s**, narrower for small values of s and wider for large values of s, the percentages remain constant for any **normal** distribution.



The point labeled **+1s** is the score one standard deviation above the mean, $\bar{X}$ + **s**; the other points are labeled accordingly. These facts are true for all **NORMAL DISTRIBUTIONS, BUT NOT FOR ALL** distributions.

Calculating a standard deviation reveals the width of the various bands, the score ranges within which the specified percentages of people score. For instance, suppose you give a newly developed test on contraception to 100 trainees. Then, suppose their mean score is 70, with a minimum score of 0 and a maximum of 100 possible on the test. If their distribution of scores has a standard deviation of 12, then you know that the distribution, if we can assume that the distribution is normal, will appear widespread and relatively flat. However, if they had the same mean score but a standard deviation of 6, the distribution would be much taller and narrower. With s = 12, 68.2% would score with 12 points on either side of the mean (34.1% above and 34.1% below) or between 58 and 82. However, with s = 6, the same 68.2% would score between 64 and 76, a much narrower range.

A word of caution is in order before discussing the practical significance of standard deviations. In distributions with a small n, even a few scores which deviate greatly from the mean can inflate the value of s and s^2, skew the distribution, and render the percentages in Figure 2, B, totally inapplicable. For example, consider the distribution of scores for 10 trainees listed below.

50 52 53 57 59 59 60 62 63 63

That distribution has a mean of 57.8 and s = 4.45 (s^2 = 19.76). However, if we lower the first score from 50 to 42, the mean becomes 57.0 but s = 6.16 (s^2 = 38.0). Consequently, you will want to look for one or two extreme scores which might indicate inferior or superior performances or attributes on the part of one or two individuals, but which do not reflect the general picture of a distribution.

So, what is the practical significance of standard deviations? How does one interpret them? Large standard deviations indicate that the group of people behind the scores is heterogeneous, varying widely with respect to the variable in question. Return to the example of 100 test scores with a mean of 70, a variance of 12, and s = 3.46. The group includes some very superior individuals and some quite inferior in their knowledge of the material covered by the test. If I had given that test as a pretest, I would be alert to the fact that I may need to accommodate beginners without boring the sophisticates. A distribution of posttest scores with s^2 = 12 would indicate that my training had failed to even out differences which

I had among trainees when I began. But I would need some pretest data, to determine how widely dispersed the trainees were at the outset, in order to make a logical interpretation. On the other hand, a small s would indicate a homogeneous group, unless I have to account for ceiling effects or basement effects (see 6.1) which could reduce s somewhat.

The values of s and s^2 become significant when you calculate inferential statistics, discussed next. Almost all parametric inferential statistics, discussed next, include the value of the variance in their calculation. In fact, many inferential parametric statistics are based upon the theory of "explaining" the variance, and the sources of it. Some of the variance arises from real differences between individuals or groups of subjects. The remainder comes from inaccuracies and errors in the measurement of the data fed into the formulae for s and s^2. For instance, a test with poor reliability creates error variance, reduces the chances of accurate results in an experiment, and inflates the variance in such a way that making evaluative interpretations of the data becomes difficult or impossible.

9.3 INFERENTIAL STATISTICS

Whenever you (1) need to compare two or more groups of individuals by analyzing data you have collected about them, (2) want to determine the degree of relationship between two variables on the same group, or (3) want to compare pretest and posttest scores, you will need to make use of **inferential statistics.** These statistics are numbers which you calculate in order to draw conclusions about the data which go beyond simply describing the data or the individuals who provided it.

Inferential statistics fall into two broad classes: **parametric and nonparametric,** described immediately below. Within each of those classes there are various kinds of statistics applied to situations depending upon (1) the nature of the data and (2) the relationships between the data and the individuals who provided them. Some are for two sets of data on the same individuals and some are for data from different groups. Others apply to one set of scores on one group.

Inferential statistics were developed and refined because scientists needed a rational, quantifiable way to determine if their observations of the world were due to random chance or due to true differences.

Suppose you give 50 trainees a pretest and they score an average of 55, then score 62 for a mean on the posttest. Did you increase their knowledge during your training or did they simply score higher on the posttest by random chance? Or, due to the reliability of the test (see 6.2)? Suppose you try two different methods of conducting new-patient instruction sessions. You want to evaluate your two methods by giving the two groups who had the two kinds of instruction a test after their different sessions. If one group scores an average of 55 and the other 65, one method seems to have a better effect; but did it really or did you simply get that difference because of chance? How much difference would you have to observe in order to confidently conclude that one method performs better than the other?

Inferential statistics were developed to answer such questions, particularly questions like the last one. They are based upon the laws of mathematics and probability.

An example will help illustrate the principles involved. Suppose you are in Las Vegas at the crap tables. You have two dice and you must roll a 7 to win $50. You let the dice fly and roll 6, a loser. The pit boss lets you take the dice home as a souvenir; after all, you should have something to take home since you have lost all of your money.

As you dejectedly pack your things and get ready to head for the airport, you decide to check the dice to see if they are fair, to see if they have been "loaded." So, you roll the two dice 36 times to see how many times 7 comes up. Of the 36 tosses, you get a 7 only once. What do you conclude? Two dice can roll to any one of 36 possible landings (six faces on each die, with 2 dice). Of those 36 possible landings, a 7 can appear in 6 ways (6,1; 1,6; 5; 5,2; 3,4; 4,3). Therefore, if your dice are fair, you should have gotten a 7 on 6 of your 36 dice rolls. But did you get a 7 once by random chance only? Or, what are the chances that two fair dice would roll a 7 only once in 36 rolls?

The laws of probability reveal that the chance of such an occurrence for two fair dice is less than 1 in 590! I would conclude that the chances were pretty good that I had been robbed!

But in statistical inferences, how much probability is necessary to make sound conclusions? If my inferential statistic has a value which indicates that I could arrive at such a value only once in 10 times by

random chance, is that sufficient? Is once in 30 sufficient? Once in 100?

If the probability is set too high, one might conclude that chance explains the difference between two groups when the two groups really do differ. If set too low, the opposite may occur; one may conclude that the data indicate a difference when in fact chance is responsible for the result.

Tradition has dictated that probability levels for inferential statistics be set at either 1 in 20 chances, typically written as .05 (1 ÷ 20 = .05) or 1 in 100 (.01).

So, when you calculate an inferential statistic, you will take your data, feed it into a formula (or the step-by-step calculation instructions provided), and arrive at a value. Next, you will compare the value of the statistic you calculated to a table of values for that statistic. The values in the table were calculated by mathematicians and were based upon the laws which apply to the probability that your statistic would have the value you observed by random chance. Typically, these tables appear in the backs of statistics books in a form such as the example below.

TABLE OF CRITICAL VALUES FOR KIRKPATRICK'S FICTITIOUS STATISTIC

n	.10	.05	.01	.001
5	3.55	4.37	5.99	83.65
10	2.73	3.23	3.67	9.80
15	1.91	2.28	2.93	3.29
20	1.58	1.81	2.02	2.25
25	1.51	1.74	1.98	2.19
50	1.46	1.61	1.78	1.95

You will find tables like the fictitious one above for any of the inferential statistics in this section in one or more of the scores of statistics books floating around. To read the table you proceed like this. Suppose you have calculated the value of the fictitious statistic and found 1.96. What does that tell you? In the column of the table labeled n, you find the n for your data, the number of data points you had. In some cases this column is labeled degrees of freedom and

equals n - 1 or n - 2.[3] (If the table does not reveal the meaning of degrees of freedom, or **df**, consult the text.) Suppose your n in this example is 20. The labels at the top of the table refer to probabilities that you would get a value of the statistic equal to the value in the table for the corresponding n.

Thus, the probability of getting a value of 1.81 for the fictitious statistic with an n = 20 is .05. Since the calculated value was 1.96, the given situation would occur by random chance in less than 5 times in 100, because the calculated value is larger than the value in the table (or the **critical value**, the value at which the calculated value becomes significant.) I would therefore conclude that I have found a true difference between the two lists of data I was comparing.

In some cases the value you calculate for a statistic must be **SMALLER THAN** the value published in the table in order to correspond to the given level of probability. The correct direction for comparison is nearly always printed near the table in statistics books.

Some statistical tables have two rows of labels for the probabilities associated with the values listed in the columns of the table. One row is labeled **one-tailed test**; the other, **two-tailed test.** These terms refer to whether or not you predict the direction of the difference between the quantities you compare via the statistical test. When you predict the direction of the difference, you are using a one-tailed test; when you do not have any basis upon which to make such a prediction, the test is two-tailed.

For instance, if you give the same group of trainees a pretest and a posttest to see if your training has taught them anything, you would expect and predict that they have learned something, and you would apply the values listed for the one-tailed test. However, if you are comparing two methods of instruction for new patients, and you have no logical reason for assuming that one group will do better on a test following the instruction than another, you apply a two-tailed test.

---

[3]In some cases degrees of freedom are not n - 1 or n - 2; none of those will appear in this manual, but they may appear in some statistical analyses, such as a Factorial Analysis of Variance, conducted by a consultant.

Thus, inferential statistics are a means of determining whether or not a given difference, or pattern of similarities, occurred because of chance or because of a true relationship (or true difference). There are numerous rules regarding which statistics you can apply to data measured at various levels (see 9.1). There are also statistics for testing similarities between two sets of scores or other data collected on the same group of individuals, called **measures of association** or **correlation coefficients.** When you want to test the difference between two (or more) means (or medians) there are tests for doing that when you have two different (independent) groups, and tests for two means based on data from the same (or from dependent) groups.

Obviously, I cannot possibly write a statistics book here. My hope is to give you enough basic information so that you can calculate some commonly useful statistics, and enough background so that you can interpret them accurately and seek help when you need it. USE CONSULTANTS!

9.3.1 PARAMETRIC VS. NONPARAMETRIC STATISTICS

When mathematicians and statisticians derive a statistical test, they sometimes make assumptions about the data to which their test will be applied. Typically, they make assumptions about the level at which the data were measured and about the shape of the distribution of the data. When they make such assumptions, they derive what are called **parametric** statistics.

When no such assumptions, or a very limited set of assumptions, apply to the derivation of the test, the statistic is called **nonparametric.** The term "parametric" refers to the parameters (mean, variance, standard deviation,...) of the population from which the samples were drawn. Sometimes nonparametrics are called **distribution free.**

Parametric statistics can be applied to those instances which meet the assumptions used when they were derived. In other instances, you can apply parametric statistics if certain conditions are met. The assumptions about the data to which parametric statistics may be applied are typically these (Siegel, 1956, p. 19):

1. The observations must be independent. That is, the subjects or cases chosen for the samples which yield the data under comparison must be chosen in such a manner that one choice does not affect the chances that another will be chosen. For instance, you make independent obeservations,

or assignments, when you assign males to one group and females to another, because the subject's sex was determined independently of your actions. But, if you pair your assignments of students to groups based upon some criteria such as age, sex, experience, education, or the like, you do not have independently assigned groups. **Another way to state the first assumption** is to say that the samples were randomly selected from the populations they represent and that they were independently assigned to the groups being compared.

2. The observations must be drawn from normally distributed populations. That means that the data must have normal distributions.

3. The data must have equal variances. The two sets (or three, or ...) of scores must have equal variances.

4. The data must be measured at the interval or ratio level. Those are pretty stiff requirements, particularly the requirement that the data be measured at the interval level or higher!

Statisticians have also proved that some of these assumptions can be violated and still permit meaningful calculations of parametric statistics. There are tests which you can use to verify whether the variances are the same or not. However, if your samples have approximately equal n's, then you can violate the assumption of equal variances. If the n's are large, then you can violate the normality-of-distributions assumption. An n is considered large when it exceeds 30, small when it is less than 20, and questionable when between 20 and 30.

The other assumptions, particularly the first one about random selection and independent assignment, cannot be violated for **parametric or nonparametric** statistics. Otherwise, the meaning of the results is opened to question.

Fortunately, when the data do not meet the assumptions outlined above, or you cannot meet the conditions allowing exceptions, one can apply nonparametric statistics to the data. Unfortunately, nonparametric statistics do not have as much **power** as parametric statistics. That means that parametric statistics will detect real differences more often when they really exist than nonparametric statistics in the same situation.

The power of a test depends on other factors, however. When you increase the n of your samples, you increase the power. Recall the fictitious table of values for the statistic, a few pages back. Suppose

that you have decided to base your probability decisions on a probability level of .05. With n = 15, you need a value of 2.28 or higher to find a significant result; with n = 20, you need only 1.81.

Thus, you can get the same power from a nonparametric test as from its parametric counterpart by increasing the size of your sample. The ratio of the size increase is called **power-efficiency** and varies from about .65 to 1.00. For example, the power-efficiency of the nonparametric Spearman's correlation coefficient is .91 compared to its parametric counterpart, the Pearson's r (both discussed below). That means that a sample .91, or 91%, as large will get a significant result for the Pearson's test in the same situation as when you apply the Spearman's test. However, if your data do not meet the assumptions of the parametric tests, then you cannot rationally apply the parametric Pearson's test. Keep this in mind when you decide which statistics you will apply to your data and when you decide on the size of your samples, if you have any choice.

Some people mistakenly scoff at nonparametrics, claiming that they are too simple and unsophisticated. But therein lies their value; you can calculate most nonparametrics with a paper and pencil in a relatively short time. Nonparametrics work well in situations when your n is small, as opposed to parametrics, which yield unstable results with small n's, particularly when n is less than 10. Nonparametrics have the added advantage of certainty when you do find significant results; anytime your nonparametric test yields a significant result, you can be assured that the analogous parametric test would also yield such a result.

Parametrics, on the other hand, are less likely to lead you to incorrect probabilistic conclusions; they have the power necessary to find significant results in borderline cases. Parametrics also enable you to apply **multivariate statistical procedures (9.3.5)**, statistical procedures which utilize large numbers of variables in the analysis and yield much more information about the data than univariate, or one-variable, tests. To make used of multivariate procedures, you will need to have access to one of the various computer programs which perform them, or to someone who does.

9.3.2 MEASURES OF ASSOCIATION (CORRELATION)

When you have two sets of data on the same group of individuals and you want to know the degree to which those two measures are alike or

different, you use **measures of association,** or **correlation coefficients.** Of the various parametric and nonparametric correlation coefficients, I will describe the Pearson's r and Spearman's Rank-Correlation Coefficient. Some texts call these statistics "measures of relationship." After describing the general ideas behind correlation, this section will discuss Pearson's and Spearman's in detail and then make some important comments about interpretations of these coefficients.

The idea of correlation. Throughout human existence, certain things seem to occur in tandem. Intelligent people seem to do well in school. Articulate people seem to sell products more easily than shy people do. People who do well in school seem to get high marks on college entrance exams. Tall people seem to weigh more on the average than short people. Of course, there are always exceptions to these general rules.

Scientists devised the correlation coefficient as a means of quantifying the degree to which two things seem to relate to each other. Consider the four parts of **Figure 3.** Each dot on the diagram represents a person. The dots' positions are determined by the individuals' scores or measured values on two variables.

In Figure 3, A, the dots form a straight line, indicating that when a person exhibits a low score or value on one quantity, then he or she is low on the other quantity, represented by the X and Y lines in the diagram. This kind of relationship represents a high, in fact perfect, correlation. In Figure 3, B, there is a lesser degree of association. Most of the time, a person who scores high on one measure scores high on the other, but there are a fair number of exceptions.

In Figure 3, C, no relationship between the two measures seems to exist. The dots appear to have randomly arranged themselves without any apparent pattern.

In Figure 3, D, the dots form a straight line once again, but now the line runs in the opposite direction from the one in Figure 3, A; when a person scores high on Y, s/he scores low on X, and vice versa.

Correlation coefficients describe in a numerical fashion the extent and direction of the relationship between the two variables. A pair of measures may correlate anywhere between a value of -1.00 and +1.00. Figure 3, A, represents two variables which have a correlation of +1.00, indicating that they vary in the same direction (high with high, low with low) and

that their relationship is perfect. Very few natural phenomena have perfect correlation, but some have correlations in the range of +.95, indicating that they are a little less than the perfect +1.00, but still quite high.

Figure 3, B, shows a situation in which the two variables have a correlation of about +.50, indicating that there is some relationship between the two, but not as strong a relationship. Correlations of .60, .70, .80, and .90 (or .68, .82, .91) indicate varying degrees between the scatter diagrams in Figure 3, A, and B. In Figure 3, C, the correlation between the two variables is .00, zero, no relationship. In Figure 3, D, the correlation is -1.00, which indicates a strong but opposite association (high with low, low with high). To test your understanding of these concepts, attempt to draw a scatter diagram representing a correlation of -.50, and then read the description of the correct shape for the scatter diagram in footnote 4, page 202.

FIGURE 3. SCATTER DIAGRAMS

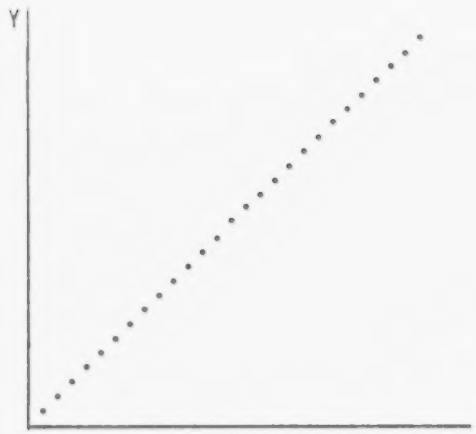


Figure 3 A
Correlation of +1.00

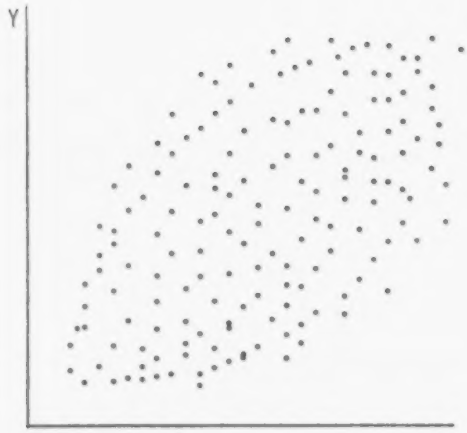


Figure 3 B
Correlation of +.50

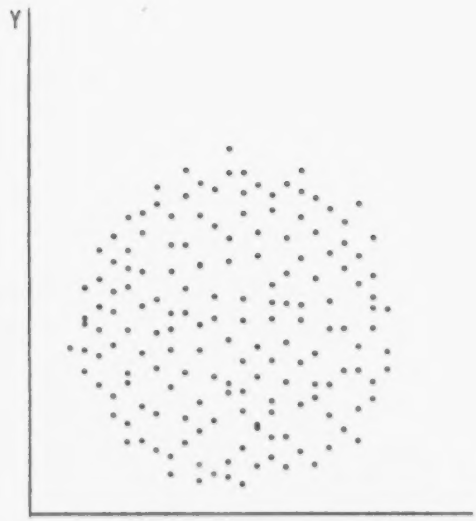


Figure 3 C
Correlation of .00 (zero)

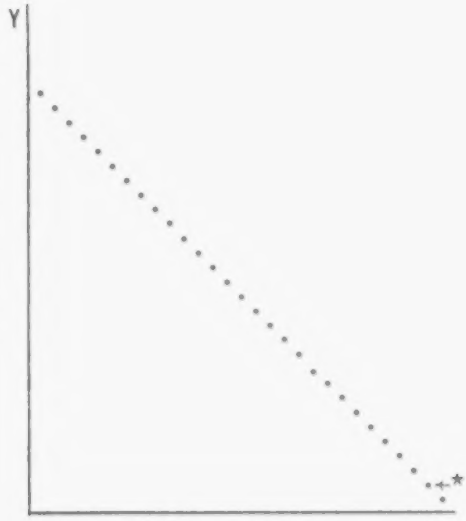


Figure 3 D
Correlation of -1.00

Each dot (.) on a scatter diagram represents an individual's score or measured value for the two variables being correlated. To make a scatter diagram, you draw two perpendicular lines, one for each of the two variables. Then, you mark off the lines into units of the two measures, such as test scores or performance ratings. Then put one dot on the graph for each person at the point corrresponding to the two scores. The dot marked with an ←* in Figure 3 D, shows a high score on the variable labeled **X** and a very low score, near 0, on Y.

Pearson's r. Usually symbolized by r or r_{xy}, this statistic is parametric. The assumptions which underlie the parametric nature of r differ somewhat from those listed in 9.3.1. Since correlatons are calculated with two measures on one group of individuals, random selection of multiple samples becomes irrelevant. However, random selection of the group from a population, for the purpose of estimating the correlation between two variables for the population as a whole, still applies.

The most significant assumption about two variables to which Pearson's r is applied is that the two variables have a linear relationship. That is, a scatter diagram of the two variables does not have a curved or hook-shaped appearance. If the scatter diagram does have such an appearance, then you cannot apply Pearson's correlation coefficient to the data and get meaningful results. To test this, make a scatter diagram of the data, making one dot for each individual at a point corresponding to the individual's scores on the two measures you want to correlate. If the scatter diagram has a curved shape, avoid Pearson's r.

The formula for calculating Pearson's r for two measures, X and Y:

$$r_{xy} = \frac{\left(\frac{1}{n} \cdot \sum_{i=1}^{n} X_i \cdot Y_i\right) - (\overline{X} \cdot \overline{Y})}{s_x \cdot s_y}$$

when n is the number of subjects for which we have values on X and Y, $\overline{X}$ and $\overline{Y}$ are the means of the two sets of values, and s_x and s_y are the standard deviations.

Calculation. To calculate a Pearson's r, follow these steps:

1. Multiply each pair of values for your subjects and find the sum of those "cross products," as they are called.
2. Calculate the means and standard deviations of both variables. (See 9.2.3 and 9.2.4.)
3. Multiply the two means.
4. Multiply the two standard deviations.
5. Divide the answer you got in step 1 by n, the number of subjects.
6. Take the answer you got from step 5 and subtract from it the product of the two means you got in step 3.

7. Take the number you got in step 6 and divide it by the product of the two standard deviations you got in step 4; the resulting number is the Pearson's correlation r for the two sets of numbers.

Spearman's Rank-Correlation Coefficient. In order to differentiate Spearman's correlations from Pearson's, most texts use r_s for Spearman's. The r_s is a nonparametric correlation coefficient based upon the same theory as the Pearson's r, and uses the relative ranks of the two measures instead of their actual values (unless the data are already ordinal ranks, in which case no conversion to ranks is necessary).

To calculate an r_s, you need to take the two sets of data for your group of individuals and **convert the raw data to ranks.** The highest value in each set of numbers gets ranked 1, the next 2, and so forth until you rank the lowest value. For instance, the data in columns A and B of the Sample Data Bank (9.2.1) would be ranked thusly. In column A, the highest value is 70 in row number 20, so that gets a rank of 1. The next highest value, 61, gets a rank of 2; 59 gets ranked 3. But then we come to four values tied at 55, in rows 14 through 17. How does the Spearman procedure require that we rank tied values? If these four values were not tied, they would be ranked 4, 5, 6, and 7; so, we assign each of the four the average of the four ranks, or the average of 4, 5, 6, and 7, which is (4 + 5 + 6 + 7) = 22/4 = 5.5. We would rank the four values of 55 at 5.5 for each of them. Proceeding similarly, we arrive at the following rankings for the two columns of data in the Sample Data Bank.[4]

---

[4]A scatter diagram for a correlation of - .50 would look like Figure 3, B, except that the points of the football-shaped area would be at the upper left and lower right.

Column A		Column B	
score	rank	score	rank
24	20	19	19
24	18	32	16
24	18	26	18
26	17	17	20
30	16	31	17
31	14	36	14
31	14	34	15
31	14	50	8.5
39	12	41	11
42	11	60	5
46	10	45	10
52	8.5	50	8.5
52	8.5	39	12
55	5.5	61	4
55	5.5	62	3
55	5.5	38	13
55	5.5	55	7
59	3	70	1
61	2	66	2
70	1	58	6

I have listed the scores in column B in the same order as they were listed in the original Sample Data Bank, but I have found that it is easier to rank a set of scores by listing them from highest to lowest on a scratch pad, ranking that scratch list, and then recording the ranks beside the scores on the original list.

That makes ranking them easier, but do not forget to keep them in the proper order, pairing the rankings of the data for each individual as in the original listing--maintaining the integrity of the pairs corresponding to each subject.

Formulas for the Spearman's r_s are derived by using the formula for Pearson's r, but applying it to ranks. This process mathematically reduces the formula to

$$r_s = 1 - \left(\frac{6 \cdot \sum_{i=1}^{n} D_i^2}{n \cdot (n^2 - 1)} \right)$$

in which n is the number of pairs (the number of subjects) and D is the difference between the ranks for each pair.

Calculation. To calcualte r_s, you can either apply the steps for calculating Pearson's r described above to the ranks, or you can follow these somewhat easier steps.

1. Make a table on a scratch pad which looks like this:

Rank X Rank Y D (Rank X - Rank Y) D^2 (D X D)

For the first two rows of the table of ranks just presented, based upon the Sample Data Bank, the table would look like this:

Rank X	Rank Y	D	D^2
20	19	1	1
18	16	2	4 (2 X 2 = 4)

Remember that if D is negative, D^2 is positive: $-2^2 = -2 \times -2 = +4$.

2. Sum the numbers in the column labeled D^2.
3. Count the number of pairs; this number is n.
4. Square n (n X n) and subtract 1 from that number.
5. Multiply the answer you got in step 4 by n.
6. Multiply the answer you got in step 2 by 6.
7. Take the result you got in step 6 and divide it by the answer you got in step 5.
8. Subtract the answer you got in step 7 from 1.00; the answer is r_s.

Interpretation of correlation coefficients. The meaning of the value which your calculations of a correlation yield varies greatly with the purpose for which you calculated the correlation. There are tables of values for correlation coefficients, based upon the number of subjects for which you have data. In the case of Spearman's r_s, the tables are arranged in terms of the n of subjects (pairs of data points); for Pearson's r the tables list n - 2 and call that "degrees of freedom." The numbers in the various columns of the tables corresponding to each level of probability refer to the probability that you would obtain a given value of r larger than .00, zero, by random chance. For example, if you have a set of 25 subjects and you calculate a one-tailed Spearman's r for that set of data which exceeds .343, then your r is significantly different from zero with a probability of .05.

But what does this mean? If you have 25 subjects and your two data points for each subject are their scores on the split-halves of a test, then your r = .343 is statistically significant, but the reliability of

your test (see 6.2.2) is far too low. If the two variables for which you collected data on your 25 subjects are their scores on a posttest you gave at the end of a teacher training program and their supervisors' ratings of their performance, then a correlation of around .35 would be considered particularly high for those kinds of variables. In fact, you might use that result as the basis for a **regression analysis,** in which you try to predict performance using the test scores.

In some cases you might choose to apply correlational techniques to research problems like the two examples in the last paragraph, or to more complex situations. If so, you would need a lot more information than I can provide here, and probably the help of a consultant.

Some people make the mistake of assuming that correlations imply causality. That is, some assume that if two variables correlate highly, then one of the two must cause the other. **Correlation should never be used to infer that one condition causes another.**

For example, assume that you are visiting the earth from another universe, gathering data about the environment, the inhabitants, the natural phenomena. You notice that each year at about the same time, in the northern hemisphere, the grass turns green and the birds migrate northward. Since those two events correlate, you might be tempted to conclude that the birds' northward migration causes the grass to turn green (or vice versa). We earthlings know that the two events occur because the seasons change in reponse to the tilt of the earth's axis and the angle at which the sun's rays strike the earth, an intervening third factor which causes both of the other two.

The same holds true for educational situations and other observed correlations; there may be another intervening factor, invisible in the data, which causes both of the effects correlated. For example, if you observe a high correlation between high school grades and college entrance exam scores, that correlation might be a result of intelligence or some other variable (or variables!).

All of this is based upon the mathematical fact that a correlation coefficient reveals how much variance two variables have in common. If you have a Pearson's r for X and Y, then the square of r, r^2, is the percentage of variance which X and Y have in common. The numbers for r and r^2 do not reveal **why** the two variances have a common percentage, only that they do.

9.3.3 COMPARISONS OF GROUPS

At times you may find it necessary to test the difference between the mean score for two groups. For example, suppose you have two groups of new patients, each of whom received instruction by different methods. You want to compare how much they learned by giving them a test of knowledge after their respective new-patient sessions. You will want to compare their mean test scores. The comparison would be performed by an inferential statistic, which, if significant, would indicate that one group scored higher than the other by more than we could expect to observe from random chance.

However, the statistical test which you would use would depend upon several different things. If (1) the test scores are interval-level data, (2) the two groups are randomly selected from your population of new patients and (3) independently assigned to the two types of instruction, then you would use the **student's t** statistic, commonly called the **t-test.** If, however, the data did not fit any one of those three conditions, then you would use a nonparametric test, probably a test called a χ^2 (read "Chi squared"; the C in chi is a hard K sound, and chi rhymes with sky). If the two samples were correlated or matched in some way, such as pairing the assignments of patients to the two types of classes by age, then you would use a dependent-samples, or correlated-samples, test, either the dependent-samples t-test or **Sandler's A** statistic.

We have twelve possible cases which might arise for selecting a statistic to analyze the difference between two groups. The two groups might be independent or correlated (matched, dependent); the data may be nominal, ordinal, or interval/ratio; and you may see instances where you can use a parametric statistic and others when you need a nonparametric test. Obviously, I cannot turn this section into a statistics text and present all of these possible tests--particularly because in some of the twelve cases there is more than one test available!

Consequently, I will refer you to any of the various statistics texts mentioned earlier in this section, to identify which test is most appropriate and to find the formula for calculating them. Please, use your available pool of consulting help, per 3.6, to identify assistants. Also, review the comments in the introductory paragraphs of this section regarding the readability of the various statistics texts which I suggested. Siegel (1956) put a very helpful table on the inside cover of his book, which classifies the various nonparametric tests case by case.

Fitz-Gibbon and Morris' (1978) text describes the calculation and correct application of the Mann-Whitney U Test, the Sign Test, the ϕ coefficient, the Point Biserial Correlation Coefficient, the Biserial Correlation Coefficient, and the t-tests for matched and unmatched samples. (Sandler's A is easier than, and equally as powerful as, the t-test for matched groups.)

9.3.4 THE CASE OF ONE OR MULTIPLE GROUPS

A whole group of both nonparametric and parametric tests has been devised for those instances when you want to analyze data from one sample or group of individuals. There is a X^2 test for the one-sample case which works for even nominal-level data.

When you have three or more groups of individuals to compare with respect to one variable (e.g. three sets of test scores for groups in an experiment) you use what is called **Analysis and Variance,** or ANOVA for short, when your data meet the assumptions for parametric tests listed above (9.3.1). While an ANOVA may be performed by hand, most social scientists perform them on a computer, with the help of a statistician who knows how to get the computer to perform the operations and how to interpret the findings.

There is an extension of the ANOVA to the instance where you have a complex set of data involving what statisticians and researchers call **factors.** Suppose you have three groups of new patients, each receiving a different kind of instruction when they enter the clinic: one group is instructed by a lecture, one by a multimedia slide/tape presentation, and one via individual sessions with clinicians. Suppose further that you suspect that age will affect how much a patient learns in new-patient class, so you further divide your three groups, for analytical purposes, into patients aged 17 and under and those aged 18 or over. Now you have six groups of patients:

	INSTRUCTIONAL METHOD		
	Lecture	Multimedia	Individual session
Aged 17 and under			
Aged 18 and older			

divided according to two factors: instructional method and age. The statistical analysis which you would use, if your data met the appropriate assumptions, would be what statisticians call a **Factorial ANOVA.**

Nothing short of a computer analysis of this kind of situation makes any rational sense; if you have ten people in each box (cell) above, then you need to perform several thousand individual arithmetic calculations in order to find the final answers. Obviously, a consultant would be needed to analyze such data, and you would be wise to enlist his or her help as early as possible in the process, to ensure that you will have analyzable data at the end.

9.3.5 MULTIVARIATE STATISTICS

The attempt to define multivariate statistics has tongue-tied more than one expert. When your analysis involves more than one outcome measure, or an attempt to predict outcome using more than one measure of past performance, you are dealing with a multivariate situation. In the example just given of a Factorial ANOVA, you would be dealing with a "univariate" statistical procedure according to some statisticians, because you have only one outcome measure, the posttest after the instruction for new patients. Clearly, however, the situation involves a multiple set of variables.

My purpose here is to do no more than acquaint you with the fact that seemingly complex situations do lend themselves to statistical analysis, although it is a gross error to assume that an appropriate statistical procedure can be identified for any situation. Unfortunately, there are many cases for which mathematicians and statisticians have not developed appropriate statistical tests.

The advantage of multivariate analyses lies in that they yield more information about the data than a set of similar but univariate tests. In fact, they control for errors which result from the repeated application of a univariate procedure to a set of outcome measures. Their disadvantages are that they are all parametric and they all require computers to calculate them; a computer can perform them in a matter of seconds, but in many cases a roomful of mathematicians would require a year or more at good calculators to arrive at the same results.

There are instances in which you cannot appropriately apply anything other than a multivariate procedure. These situations almost always arise

when you conduct experimental research, but they can appear when you want to analyze data which already exist in some form, data which you want to analyze for the purposes of a needs assessment, for example.

Suppose the medical records in your clinic contain two kinds of pill patients: those who continue using the pill for a year or more and those who discontinue use for some reason other than pregnancy in that year. From all those patients, you obtain data such as their age, their number of years of contraceptive use, and their number of years of sexual experience. There is a procedure called a **discriminant function analysis** which would tell you which of those three variables is most likely to predict whether future patients will fall into the continued-users group or the group which discontinues use. The analysis may show that not one of the three variables predicts anything you are interested in knowing. But it may reveal a great deal about how you should educate new pill-patients.

Summary. I hope that this brief and incomplete summary of the various statistical procedures will prove useful in a vast majority of cases and give you enough information to know where to seek help and to know when you have reached your limitations.

I believe that we family planning educators have learned far too little about what we do, in part because we have not taken advantage of the tools available to us. If we can rely upon consultants who have a familiarity with those tools, then we can begin to document and compile a body of knowledge about our profession that goes far beyond what now exists and, perhaps, beyond even our own hopes.

SELF-TEST ON SECTION NINE

You will probably get more information about how much of this material you know if you answer the questions below on a separate sheet of paper and then compare your answers to the ones given on the following pages.

1. What are levels of measurement? Why is it important to distinguish among them?

2. Define nominal, ordinal, interval, and ratio levels of measurement.

3. What level of measurement does each of the following measures attain?

   a. the balance in your checkbook
   b. zip codes
   c. experience on the job, expressed in years, months, and days
   d. the use-effectiveness of various contraceptives
   e. ranking of trainees at the end of a course or workshop

4. What are the advantages and disadvantages of a frequency distribution versus a grouped frequency distribution?

5. When constructing bar graphs, how high should the tallest (or how long should the longest) bar in the graph be in relation to the size of the boundaries of the entire graph?

6. How much is 42% of 237?

7. What percentage of 8,100 is 2,500?

8. If you trained 327 professionals last year and you trained 396 this year, what was the percentage of change?

9. Using the data from column B of the Sample Data Bank (9.2.1), find the mean, median, and mode of that distribution.

10. What is skewness? Why would you want to know that the distribution of test scores for a pretest is negatively skewed?

11. Find the range, variance, and standard deviation of the data in column B of the Sample Data Bank (9.2.1).

12. If a distribution is skewed, what conclusion can you draw about the percentage of the distribution which lies within one standard deviation of the mean?

13. Of the various assumptions for applying a parametric statistical test, what is the one assumption which also applies to nonparametrics?

14. Calculate Pearson's r and Spearman's r_s for the data in the Sample Data Bank (9.2.1). What circumstances would dictate that you report Pearson's correlation coefficient and which would require the Spearman's coefficient if these data were real?

15. Suppose the data in the Sample Data Bank are posttest scores from a workshop you conducted (column A) and supervisors' ratings of the traineees after they have been back on the job for a month (column B). Given the correlations you calculated in question 13 above, which would you report? Could you conclude that your workshop caused the kinds of ratings the trainees earned from their supervisors?

16. What is an independent samples t-test? When can you apply it?

ANSWERS TO SELF-TEST ON SECTION NINE

1. Level of measurement refers to the degree of fineness to which one measures or quantifies a variable. The distinctions between them are important when deciding which descriptive and inferential statistics are appropriate for any kind of data.

2. The definitions of the four levels of measurement appear in 9.1.

3. a. ratio
   b. nominal
   c. ratio
   d. at least interval level (ratio if zero means "no effectiveness")
   e. ordinal

4. A grouped frequency distribution requires less space and more clearly identifies the number of persons or objects which fall into categories of interest, such as a breakdown by age, prior performance, or the like. A frequency distribution takes more space but allows the reader (or the original writer who does not have the raw data any longer) to calculate descriptive statistics from the data presented; a grouped frequency distribution does not.

5. In order to avoid overemphasis of the largest bars, they should not exceed 80 to 90 percent of the total space allotted for bars. See 9.2.1 for more details.

6. $\frac{42\%}{100} = .42$; $.42 \times 237 = 99.54$

7. $\frac{2,500}{8,100} = .3086$; $.3086 \times 100 = 30.86\%$, or about 31%

8. $\frac{396 - 327}{327} = +69 = +.211$; $+.211 \times 100 = +21.1\%$, a 21% increase.

9. Mean = 44.5; median = 43; mode = 50.

10. Skewness refers to a distribution which gives the general appearence of a normal distribution but which "leans" to one side or the other. A negatively skewed distribution has the mass of scores toward the upper values, with a smaller number strung out along the lower values. If you had a distribution of pretest scores which was negatively skewed, then you would know that the mass of your trainees scored highly on the pretest but that there are a minority who scored low. Thus, without boring the majority who scored highly on the pretest, you would need to consider an attempt to bring the ones who scored lower up to the same entry level, so as not to make too many assumptions about what they know at the outset of the session(s).

11. Range is 19 to 58; variance = 231.95; standard deviation = 15.23

12. You can draw no conclusion at all. The figures for the percentages of a distribution which fall within the bands divided by the standard

deviations above and below the mean in Figure 2 of this section apply to normal distributions only; a skewed distribution is not normal.

13. Nonparametric tests, as with any inferential statistical procedure which attempts to identify differences between samples or to define the parameters of a population based upon the observed values of those parameters for a sample, assume that the sample or samples were randomly selected for inclusion in the groups and independently assigned to the groups. The assumptions which the nonparametrics do not have to follow are those regarding the variance and normality of the distribution.

14. $r = .81$; $r_s = .85$. If the data are normally distributed, or the variables which yielded the data can be assumed to be normally distributed, if the data are at least interval-level measures, and if the scatter diagram of the data is not curved or hook-shaped, then you could report the Pearson's r. If any one of those conditions is not met, then the Spearman's correlation is appropriate.

15. You would report the Spearman's r_s because the supervisors' ratings of job performance are oridinal data, thus eliminating the Pearson's r because of the requirement that a Pearson's r be calculated for data measured at least at the interval level. You could **not** conclude that your workshop increased supervisors' ratings; correlation does not imply a cause-and-effect relationship between variables or quantities. In this case, a long list of other factors besides your workshop could affect the supervisors' ratings.

16. The independent samples t-test is a parametric inferential statistic for comparing the difference between means of two samples to determine if the difference between the means is significant or if it appeared in the data by random chance. You can apply the t-test when your data meet all of the criteria listed in 9.3.1.

section ten:
A THOROUGH EXAMPLE

In this section I propose three priorities for a hypothetical affiliate of the Planned Parenthood Federation of America to use in planning its education program. Then, two goals for the highest priority--teenagers--are drafted, along with objectives for the goals. Next, I follow these objectives through the remainder of the planning and evaluation process described in **section three.**

This section contains:

1. A discussion of how I decided upon the various standards which the objectives set for the education program to accomplish.

2. Strategies for collecting the data to evaluate against the objectives.

3. A sample task analysis for a workshop conducted for school nurses.

4. Comments about how I might interpret the evaluative data and about its implications for future program planning.

5. Four sample questionnaire instruments: a high school speaking engagement feedback form for teachers, a sample workshop needs assessment questionnaire, a content and process evaluation form for a workshop, and a questionnaire to distribute to new patients to determine how they learned about the affiliate's services.

All of the people, places, and events in this section are figments of my imagination. Any resemblance to any affiliate, any location, or any person, living or dead, is entirely coincidental. While the example presented is realistic, **it should not be construed as a prescription** for any particular community or family planning program.

There is no self-test at the end of this section.

section ten: A THOROUGH EXAMPLE

> "Few things are harder to put up with than the annoyance of a good example."
>
> --Mark Twain

Using the rest of this manual as a background, I will now present a thorough example of how the educators at a hypothetical affiliate of the Planned Parenthood Federation of America might make use of this manual. Since a complete example covering a year's worth of programs would require another book the length of this manual, I will analyze the evaluation and planning processes for two specific objectives only. However, I will begin with the statement of priorities for the education program, write examples of goals for the first priority, and then write examples of objectives for those goals.

10.1 BACKGROUND INFORMATION

For purposes of this example, I will assume that I am the education director at Planned Parenthood of Midwest City, a fictitious locale somewhere in the United States. In addition to myself, there are two other educators--a community health educator and a trainer--plus a secretary, working in the education department. The population of 300,000 in Midwest City includes 20 percent (60,000) blacks, 8 percent (24,000) Hispanics, and the remainder approximately half Protestant and half Catholic "Anglo" Caucasians.

Polls conducted by local news organizations indicate that Midwest City fairly typically represents the nationwide opinions on sex education, the right to choose, and related issues regarding reproductive freedom. About eighty percent of the adults in the community favor sex education which includes instruction in contraception for students in junior and senior high schools, while only four percent are opposed to that. Less than five percent oppose legal abortion, including the membership of a very active and vocal anti-choice group. A similarly small conservative group, one percent of the population of Midwest City, vocally opposes any form of sexuality education in the public schools.

Planned Parenthood of Midwest City serves an active patient load of 10,000, with approximately 6,500 patient visits annually. The affiliate operates one centrally located clinic in the downtown area, housed in two buildings. The education and counseling departments have their offices in

an old mansion which has been subdivided into office space.

Of the annual budget of approximately $600,000, $65,000 is spent on education.

Of that total for education, $38,000 goes to salaries and fringe benefits, $15,000 for operating expenses (rent, utilities, telephone, printing, postage, insurance, local travel, and office supplies). The remaining $12,000 is spent for equipment (projectors, typewriters, adding machines, and the like), films and educational materials (books, literature...). Any out-of-town traveling done by education staff members must come out of the $12,000 budget for educational materials.

10.1.1 SETTING PRIORITIES

Data collected from various sources indicate the following salient facts about Midwest City and its people.

1. The birthrate for households with total incomes below the federal poverty level is 29.5 births per 1,000 population, or 90 percent larger than the birthrate of 15.5 for Midwest City as a whole.
2. The state department of human services reports that 52 percent of the people receiving Aid to Dependent Children (ADC) are households headed by single mothers aged 18 through 24.
3. The birthrate for single teenagers rose 2 percent per year throughout the 1970's to its current level of 30 births per 1,000 teenagers in Midwest City.

No other organization or agency in Midwest City other than the private medical community provides family planning services. The local health department provides free treatment for sexually transmissible diseases (STDs) to anyone who has reason to suspect that they have contracted an STD, without regard for income level or residency.

Based upon those data, I would set three priorities: (1) teenagers, (2) low income and minority groups, and (3) welfare recipients. I would assume that those three priorities imply a fourth priority: that any educational resources remaining after efforts have been aimed at those three groups would be aimed at individuals not included in those first three priorites.

10.1.2 GOALS

Based upon the needs assessment data and the priorities I have set, I would write goals for each priority. I will, however, include only the goals which I have in mind for the first priority, teenagers:

Goal 1: We will increase use of reliable contraceptives among sexually active teenagers in Midwest City.

Goal 2: We will increase the numbers of new, sexually active teenage patients at Planned Parenthood of Midwest City.

10.2 PLANNING THE EDUCATION PROGRAM

At this point I will discuss how I would write behavioral objectives and perform the other tasks in steps 2 through 5 of section three, the Planning Phase.

Behavioral objectives. I would write the following behavioral objectives aimed at accomplishing the two goals. Since the two goals are similar in their intent, and might, in fact, be stated as a single goal, I will write only one set of objectives. I wrote two goals for reasons which will become obvious as I list my objectives. I simply want to shorten the process of referring back to my objectives by having one list with consecutive numbers.

1. During 1981 the community health educator and the education director will speak to a total of 1,750 teenagers in schools and at other locations, an increase from 1,500 reached in 1980.

2. During 1981 the education director and the trainer will conduct three workshops, of four hours' duration each, for school nurses from the three school districts in Midwest City, to train the nurses to make knowledgeable referrals of students to Planned Parenthood of Midwest City.

3. During 1981 the education staff will secure air time for 500 public service announcements aimed at recruiting new teenage patients over local radio stations which aim their programming at people aged 12 to 25.

4. During 1981 the education staff will recruit 1,000 new teenage patients who will seek contraceptive services at Planned Parenthood of Midwest City.

5. At the end of 1981 the education director will obtain the necessary vital-statistics data to determine the percentage of change in the birthrate for single teenage mothers from 1980 to 1981 and write a report on the findings.

6. During 1981 the education director will conduct a survey of new patients during three different time periods (in April, September, and December) in order to determine the percentages of new patients recruited through speaking engagements, referrals by school nurses, radio PSAs, and word of mouth between old and new patients.

How did I determine the standards which those objectives set? The reasoning behind the numbers of teenagers to be contacted in schools and other locations in objective number 1 will become clearer when I perform the task analyses for that objective. I set that standard based upon the amount of time which the community health educator will have available during 1981 for doing speaking engagements. I set a goal of 500 public service announcements in objective number 3 because my experience in working with radio stations and PSAs has been that each station will air an announcement about once every three days. I simply assumed that there were therefore 100 opportunities in the year for each radio station and assumed that there were five radio stations aimed at teenagers in Midwest City. I set a standard of three workshops for school nurses in objective 2 because there are three school districts in Midwest City, and I have found that it is logistically easier to conduct such workshops one district at a time.

How did I determine an objective of 1,000 new teenage patients in objective number 4? I have no rational reason to believe that 1,000 is either too large or too small. I simply set the standard at which to aim while I gather background data. Next year, for 1982, I will have more information about how many patients the education programs can recruit. Objective number 5 is aimed at determining whether or not my education program is moving toward goal number 1. Objective 6, to survey new patients regarding how they were recruited, will combine with the data in number 5 to help me determine two important facts: Which method of recruiting new patients brings in the largest number of new patients? When compared to the data gathered to accomplish objective number 5, do the efforts outlined in objectives 1, 2, and 3 seem to be moving me closer to my goals?

I need to ask myself these questions in order to fulfill step 3 in the planning and evaluation process. That step requires me to ascertain if accomplishing my objectives will move me closer to my goal.

Data collection and task analyses. Collecting the data necessary to verify the achievement of objectives 1 through 4 will be a simple matter of maintaining good records. I will assume that the data necessary to accomplish objective 5 will be available from the state health department, the county health department, or some other organization responsible for collecting vital statistics. I am safe in assuming that because every

state does so. The data necessary to accomplish objective number 6 will become available when the education director conducts the surveys required in the objective.

My task analyses for objective number 1 lead me to the assumption that the community health educator and the education director could reach 1,750 teenagers during 1981. Based upon my objectives for accomplishing other goals, I know that the education director and the health education specialist will have other responsibilities during the year. The health educator will have to teach new-patient classes, coordinate the film lending library, and make other kinds of presentations during the year. The education director has other responsibilities in the objectives listed here, as well as responsibilities in supervision, report writing, training, purchasing, and so forth. Based on my analyses of their other tasks, I know that the education director and the community health educator have approximately 250 hours during the year to devote to speaking engagements. Each speaking engagement represents approximately two and one-half hours of time if I include phone calls, travel, and time to make the presentation. Therefore, I will have time for the education director and health educator to do a total of 100 speaking engagements (250 hours divided by 2.5 hours per speech). From my records for 1980 I know that a typical speaking engagement has an average teenage audience of 17.5 participants. (That is, the mean number of people in my speaking-engagement audiences during 1980 was 17.5) Thus, 100 speaking engagements with an average audience of 17.5 people gives me a total of 1,750 as a reasonable goal based upon my allocation of time and other resources.

The following is a more detailed task analysis of the process involved in conducting ONE of three workshops for school nurses. I am assuming in the following task analysis that I am analyzing the first of the three workshops, with the hope that the remaining workshops will take slightly less time.

SAMPLE TASK ANALYSIS:

A Four-Hour Workshop for School Nurses

Task		Number of hours			
		Sect'y	Director	Trainer	Total
1.	Contact and meet with school district director of nurses.	0	2	2.5	4.5
2.	Create, revise, prepare needs assessment questionnaire	1	1	2	4
3.	Mail, receive returns of quest'aire	2	0	0	2
4.	Compile results of needs assessment and set workshop agenda	0	2	3	5
5.	Create flyer announcing workshop	1	0.5	1.5	3
6.	Mail flyer, receive registration forms	2	0	1	3
7.	Prepare workshop agenda for distribution to participants, make copies	1	0.5	1	2.5
8.	Select literature for nurses to receive, prepare packets, set up meeting room	3	3	3	9
9.	Create content and process evaluation questionnaire for workshop, make copies	1	0.5	1.5	3
10.	Conduct workshop	0	5.5	5.5	11
11.	Clean up meeting room	0.5	0.5	0.5	1.5
12.	Compile results of evaluation questionnaire, write report	1	2.5	3	6.5
13.	Write thank you notes to guest speakers (if any), answer phone calls from participants before and after workshop	2	3	3	8
	Totals	14.5	21	27.5	63

Subsequent workshops for nurses would elminate only task numbers 2 and 9, thus reducing the total number of hours required to conduct the workshop by only about 7 of the total of 63 hours. Therefore, the accomplishment of the objective which requires the educators to conduct three such workshops would require a total of 175 working hours (63 for the first one plus 56 for each of two more).

With the above task analysis, you could estimate the total cost of conducting the workshop by estimating the cost for those items described in **Figure 3** of **section three** (3.3.4).

10.3 IMPLEMENTATION

With written behavioral objectives and task analysis for each objective, I would have a clear and specific knowledge of what the other education staff members and I would have to do during 1981. This would enable us to do time management analyses and accomplish our tasks in order and on time.

In section 10.5 I have included sample forms for the needs assessment questionnaire for the nurses' workshop, a content and process evaluation questionnaire for the nurses' workshop, and the survey questionnaire for new patients as required in objective number 6.

Since the work of family planning educators never quite goes as planned, I want to include one other sample form in section 10.5. Since Midwest City has an active anti-choice and anti-sex-education group, I want to make sure, as education director, that I can document what happens each time I go to speak to a group of teenagers. Therefore, as promised in an earlier section, I am including in 10.5 a sample questionnaire for high school teachers and other teenage-group leaders to complete after speaking presentations which my educators or I give during 1981.

10.4 EVALUATION

If the evaluative data is collected as required in objectives 1 through 4, one can immediately see whether those objectives have been met or not. At the end of 1981, after collecting data to indicate how many new teenage patients came to Planned Parenthood of Midwest City, I will have valid background data for setting a more reasonable standard for 1982.

The data required by objectives 5 and 6 would require some additional interpretation, however. As was indicated in **section four**, trying to draw cause-and-effect conclusions from vital-statistics data, such as that collected for objective number 5, is frought with difficulties.

If the number of teenage pregnancies continues to increase, that would not **necessarily** indicate that the rate of teenage pregnancy, or the birth-rate for single teenage mothers, has increased. The number of teenage pregnancies might increase because the population of teenagers in Midwest City increased. The number of teenagers living in Midwest City might increase because of migration into the area or because more 12-year-olds became teenagers at their thirteenth birthday than 19-year-olds reached their twentieth birthday during 1981. You may want to refer back to the discussion in **section four** on these issues.

However, if the birthrate (numbers of births to teenage mothers per 1,000 teenagers in the population) changed during 1981, then I would want to look for possible explanations. I would be particularly interested in either a decline in the birthrate for teenage mothers or in an increase in the birthrate smaller than the increases seen over the preceding few

years. I would feel professionally safe and ethical in claiming some responsibility for a decrease, or less of an increase, if I could document that my agency's efforts have brought more sexually active teenagers into the ranks of those using reliable contraception.

If, on the other hand, I saw an increase in the birthrate for single teenage mothers during 1981, I would be concerned about how my education program for 1982 might be changed to reverse that trend. If the birthrate increased and I could document that I had recruited several hundred new teenage contraceptive patients, then I would have to ask myself why the birthrate increased. There are several possible explanations. Perhaps more teenagers were sexually active in 1981 than in previous years. Perhaps the new teenage patients I recruited obtained a reliable method of contraception and then failed to use it. Perhaps the number of new teenage patients I recruited was too small to have a marked effect. Whatever the explanation, if I saw an increase in the birthrate for single teenagers, then I would want to attempt to discover the cause during 1981, so that I could more effectively apply my educational resources to the problem.

The data collected from my survey (objective 6) should reveal a lot about the relative effectiveness of various means of recruiting new teenage patients. I might discover that one method of recruiting patients worked much better than another, or that they work with about equal effectiveness.

By combining the data collected in objective 4 with the results of the survey for objective 6, I would be able to make estimates of the cost to recruit patients by the various techniques. For instance, suppose that I discovered that exactly 1,000 new teenage patients came to Planned Parenthood of Midwest City seeking contraception in 1981.

Suppose also that my survey of new patients found that 60 percent of all new teenage patients came to the clinic as a consequence of a radio PSA, high school speech, a referral by a nurse, or something else which the education department did during 1981. From those facts, I would know that 60 percent of the 1,000 new teenage patients, or 600, came to the clinic because of something the education department did. Suppose further that my survey results indicated that one half of those 600 new patients the staff recruited came because of a speaking engagement. From my task and cost analyses, I could determine how much money, time, and other resources were involved in recruiting one of those 300 new patients. Similarly, I could

determine the cost of recruiting one new patient through radio PSAs, through training school nurses, or any other method I happen to use.

That information may not necessarily dictate that I put all of my resources for recruiting new patients into one method of patient recruitment. However, it might indicate that the more costly methods are recruiting relatively few patients at a higher price per new patient recruited.

Summary. I hope that the information contained in this section has been instructive and illustrative of some of the material which has preceded it. More than that, I hope that this section has helped to inspire your thinking in terms of how planning and evaluation can simplify your work as a family planning educator.

10.5 SAMPLE FORMS

I have included four sample forms in this subsection:

Sample Form:	Content:
A	High School Speaking Engagement Evaluation Form
B	Needs Assessment Questionnaire for Nurses' Workshop
C	Content and Process Evaluation Form for Nurses' Workshop
D	New Patient Recruitment Survey Form

I would probaby not label the forms by the above titles when I distributed them to my participants or subjects; I have typed each form with the above labels (in **bold face**) for the purposes of this manual only.

These samples are intended as examples, not as universally applicable instruments. If you find that one or more of them are useful to you, then feel free to prepare them for distribution or to make copies of them directly from these pages, after removing the headings of this manual. However, our typing requirements for this manual may make some of the forms longer than necessary. If you can use them, try to put them onto one page; that will increase your response rate. People seem less intimidated by one page than by two or more.

10.5.1 SAMPLE FORM A: HIGH SCHOOL SPEAKING ENGAGEMENT EVALUATION FORM

HIGH SCHOOL TEACHERS' FEEDBACK FORM

Planned Parenthood of Midwest City rev. 1981

We need your help! When our speakers represent us in the community, we like to get feedback on their performance from those who invited us to speak. Please take a few minutes NEAR THE END OF THE SPEECH OR AFTER IT IS OVER to answer these questions.

Please answer the following questions by writing YES or NO in the space before each one. DID THE SPEAKER...

_____1. Arrive on time?
_____2. Cover the material which you requested to be covered?
_____3. Maintain the audience's attention throughout the speech?
_____4. Seem to know his/her material as well as you had hoped?
_____5. Speak so as to be heard?
_____6. Get any questions from members of the audience AFTER the talk?

7. Were there any topics which you asked your speaker to avoid during the presentation? If so, list them below and state whether you gave the speaker permission to answer students' questions about the topic or whether you asked that that topic be totally avoided.

Topics	Questions OK	Asked to avoid
_____	_____	_____
_____	_____	_____
_____	_____	_____
_____	_____	_____

8. Did any of the following topics arise during the speaking presentation? If so, please, put a check mark beside the topic indicating whether the speaker raised the topic, a student raised it, or whether you, the teacher, asked about it.

TOPICS	Was not discussed	Topic was raised by: PP speaker	student	teacher
A. Abortion	___	___	___	___
B. Homosexuality	___	___	___	___
C. Masturbation	___	___	___	___
D. Oral sex	___	___	___	___
E. Premarital sex	___	___	___	___
F. Parental consent for a minor to get birth control	___	___	___	___

9. Please use the back of this form to write down what you recall our speaker saying about any of the topics listed in question 8. Please identify the topic to which you are referring.

10. Please use the back of this sheet or additional sheets if needed to tell us anything you would like us to know about the speaker or the presentation in your class today. No one other than PP Midwest City staff and board members will see this form unless we have your permission first.

_____ _____

Your name School

_____ _____

Today's date Class period for which this form completed

THANK YOU!

10.5.2 SAMPLE FORM B: NEEDS ASSESSMENT QUESTIONNAIRE FOR NURSES' WORKSHOP

SCHOOL NURSES' WORKSHOP
PLANNING QUESTIONNAIRE

The Education Department of Planned Parenthood of Midwest City is going to conduct a workshop on referring students for family planning services, in cooperation with the J. F. Kennedy Independent School District. In order to ensure that we offer what you want and need, we would like you to answer the following questions and return the form to the address given at the end BY FRIDAY, MARCH 6, 1981. No one other than PP Midwest City staff will see the questionnaires.

10.5.2

1. We need to make sure that the topics we cover will be useful to those who attend. Please read the list of topics on the left and check the appropriate space on the right beside each topic.

TOPICS:	I FEEL LIKE I: Need in-depth coverage	Need a brief review	Am already knowledgeable
A. Planned Parenthood services for teenagers	____	____	____
B. Teenagers' legal rights with respect to sexuality	____	____	____
C. Assisting teens with decisions	____	____	____
D. Your district's policies on interviews, confidentiality, referrals	____	____	____
E. Homosexuality & teenagers	____	____	____
F. Practice interviews/role plays	____	____	____
G. Other topics (please specify)_____			

_____

_____

2. Would you be interested in registering for continuing education credits from the State Board of Examiners for Nurses? ___Yes ___No

3. How much is the largest fee you would be able to pay, if one is charged, to attend this workshop?
____none _____up to $5 _____up to $10 _____up to $15

4. We need to plan the workshop at a time when the maximum number of nurses can and will attend. Please put a checkmark beside the times listed below WHICH WOULD FIT WITH YOUR SCHEDULE:

_____Monday afternoon from 3:30 to 7:30

_____Tuesday afternoon from 3:30 to 7:30

_____Wednesday afternoon from 3:30 to 7:30

_____Thursday afternoon from 3:30 to 7:30

_____Tuesday evening from 6:00 to 10:00

_____Thursday evening from 6:00 to 10:00

_____Saturday morning from 8:30 a.m. to 12:30 p.m.

Please return this form by Friday, March 6, 1981, to:
Education Department
Planned Parenthood of Midwest City
2203 Roosevelt Ave.
Midwest City, BT 45678

10.5.3 SAMPLE FORM C: CONTENT AND PROCESS QUESTIONNAIRE, NURSES' WORKSHOP

Note. The responses to the preceding sample questionnaire would help me determine the agenda for the workshop. From that agenda I would compose the list of topics in the first items of the sample questionnaire below.

PARTICIPANTS' FEEDBACK FORM

School Nurses' Workshop

April 4, 1981 Planned Parenthood of MWC

Since we will be doing this workshop again for other districts' nurses, we would appreciate it if you would take a few minutes and complete the questions below, so that we can make any necessary changes in our workshop.

1. If you were going to attend this workshop FOR THE FIRST TIME, would you have us change the agenda? Please indicate whether you would have us spend more time, less time, the same amount of time, or leave each of the following topics completely out of the next workshop, by putting a checkmark in the appropriate column.

TOPICS	More of it	Less of it	The same	Leave it out
A. **(List the topics covered here,**	___	___	___	___
B. **in chronological order)**	___	___	___	___
C.				

Please indicate the degree to which you agree or disagree with each of the following statements by circling the appropriate response to each statement in the list below:

2. I will feel comfortable about referring students to Planned Parenthood of Midwest City.

Strongly agree Agree Disagree Strongly disagree

3. What I learned in this workshop will be helpful to me on my job.

Strongly agree Agree Disagree Strongly disagree

4. The workshop leaders know the material they covered.

Strongly agree Agree Disagree Strongly disagree

5. Did you think that this workshop had the right balance of lectures and exercises? Please circle one number on the line below to indicate your opinion. Circle 1 to indicate that there was way too much lecture, 7 to indicate that there were way too many exercises, and 4 to indicate that there was the right balance. If your opinion is somewhere between, circle one of the numbers between the middle and the end.

1 2 3 4 5 6 7

Too much lecture — Right balance — Too many exercises

6. Please use the space below and the back of this sheet to tell us anything else you would like us to know about this workshop.

THANK YOU!

10.5.4 SAMPLE FORM D: NEW PATIENT RECRUITMENT SURVEY QUESTIONNAIRE

NEW PATIENTS:

WE NEED YOUR HELP! We want to let everyone who needs us know about our clinic. Will you help us by taking a few minutes to answer some questions? DO NOT PUT YOUR NAME ON THIS PAPER. No one except PP staff will know what answers you give.

1. Where did you first learn about Planned Parenthood of Midwest City?

_____Newspaper
_____Radio
_____TV
_____Teacher/counselor
_____School nurse
_____Doctor/other health agency
_____Planned Parenthood speaker
_____Friend
_____Another Planned Parenthood patient
_____Parent/brother/sister
_____A friend who is not a patient here
_____Other_____

2. If you first heard about Planned Parenthood from a FRIEND, RELATIVE, OR ANOTHER PLANNED PARENTHOOD PATIENT, do you know how that person heard about us?

_____Newspaper
_____Radio
_____TV
_____Teacher/counselor
_____Planned Parenthood speaker
_____Another Planned Parenthood patient
_____Parent/brother/sister
_____School nurse

What is today's date? _____
What is the zip code at your home? _____
How old are you?_____ Are you _____ male or _____ female?
Do you consider yourself to be a
_____Black
_____Native American Indian
_____Spanish-Speaking American (Chicano, Puerto Rican, Cuban, Other)
_____Caucasian
_____Asian or Pacific Rim ancestry

THANK YOU!
Please give this form to the person teaching your new-patient class.

section eleven:
SELECTED BIBLIOGRAPHY

This bibliography contains all of the citations in the text, along with some additional references.

A selected list of centers that may provide you with additional information appears in section three. Section four contains a list of federal publications which include potentially useful data. Three sources of more information about tests appear at the end of section six.

Rating system. I have attempted a rating system for the materials listed in terms of their **level of difficulty:**

* = **elementary;** the material presumes no background in the subject and no fundamental knowledge beyond what would be offered in a high school curriculum with a college preparatory emphasis.

** = **beginners;** the material presumes an elementary knowledge of the material, such as that offered in a beginning undergraduate course on the subject. (In the case of a statistics text or article, it presumes a basic knowledge of terms and an understanding of arithmetic and high school algebra.)

*** = **intermediate;** presumes knowledge included in materials rated * and ** and usually indicates a text which is used at the undergraduate level. (In the case of statistical works, presumes a knowledge of college algebra and perhaps some matrix algebra.)

**** = **advanced;** presumes a thorough understanding of the field and of related areas; reserved almost exclusively for texts and materials used at the very advanced undergraduate level or the graduate level.

The ratings do **not** indicate the quality of the material nor are they an endorsement of the author, publisher, copyright holder, or content. The rating system is intended to guide the user's selection of materials so as to avoid the frustrations of ordering or searching for a work and finding it too advanced. The absence of a rating does not signify anything other than my decision that the material did not seem to lend itself to the rating system.

Alan Guttmacher Institute. 11 Million Teenagers. New York: Planned Parenthood.

** Babbie, Earl. Survey Research Methods. Belmont, CA.: Wadsworth, 1973.

*** Bennett, C. A. & Lumsdane, A. A. (Eds.). Evaluation and Experiment. New York: Academic, 1975.

**** Bohrenstedt, G. W. "Reliability and Validity Assessment in Attitude Measurement." In G. F. Summers (Ed.), Attitude Measurement. Chicago: Rand-McNally, 1970, pp. 80-89.

**** Brach, Glenn H. & Glass, Gene V. "The External Validity of Experiments." The American Education Research Journal, 1968, vol. 5, pp. 437-467.

Buros, O. K. Mental Measurement Yearbook (8th. ed.). Edison, NJ: Gryphon, 1978.

*** Campbell, Donald T. & Stanley, Julian C. Experimental and Quasi-Experimental Research Designs in Education. Chicago: Rand-McNally, 1963.

** Caro, F. G. (Ed.). Readings in Evaluation Research. New York: Russell Sage Foundation (230 Park Av. 10017), 1975. (ERIC No. ED060929)

Cooley, W. H. & Lohnes, P. R. Evaluation Research in Education: Theory, Principles, Practice. New York: Irvington, 1976

*** Cromback, L. J. Essentials of Psychological Testing. New York: Harper, 1960.

** Fox, D. J. The Research Process in Education. San Francisco: Holt, Rinehart, and Winston, 1969.

**** Glass, G. V. & Stanley, J. C. Statistical Methods in Education and Psychology. Englewood Cliffs, NJ: Prentice-Hall, 1970.

*** Goldman, L. Using Tests in Counseling (2nd ed.). New York: Appleton, Century, Crofts, 1971.

*** Goldman, L. (Ed.). Research Methods for Counselors. New York: John Wiley and Sons, 1978.

*** Guttentag, M. & Struening, E. L. Handbook of Evaluation Research. Beverly Hills, CA: Sage, 1975.

Harless, J. H. An Ounce of Analysis (Is Worth a Pound of Objectives). Newman, GA (218 Jackson St.): Harless Performance Guild, Inc., 1975.

**** Harris, R. J. A Primer of Multivariate Statistics. New York: Academic, 1975.

BIBLIOGRAPHY

Haveman, R. H. "Benefit-Cost Analysis and Family Planning Programs." Population and Development Review, 1976, vol. 2(1), pp. 37-64.

** House, E. R. Evaluating with Validity. Beverly Hills, CA: Sage Publications, 1980.

* Huck, Schuyler W., Cormier, William H., & Bounds, William G., Jr. Reading Statistics and Research. New York: Harper & Row, 1974.

Kantner, John F. & Zelnik, Melvin. "Sexual Experience of Young Unmarried Women in the United States." Family Planning Perspectives, 1972, vol. 4 (4), pp. 9-18.

**** Kerlinger, Fred N. Foundations of Behavioral Research (2nd ed.). New York: Holt, Reinhart, and Winston, 1975.

**** Kerlinger, F. N. & Pedhauzer, E. J. Multiple Regression in Behavioral Research. New York: Holt, Rinehart, and Winston, 1973.

** Kirkpatrick, J. S. "Nonparametric Statistics: Useful Tools for Counselors." The Personnel and Guidance Journal, 1981, vol. 59(9), in press.

Lief, Harold I. & Reed, David M. Sex Knowledge and Attitude Test. Philadelphia: University of Pennsylvania, 1972.

Lord, F. M. "The Relationship of the Reliability of Multiple-Choice Tests to the Distribution of Item Difficulties." Psychometrica, 1952, vol. 17, pp. 181-194.

* Mager, R. F. Goal Analysis. Belmont, CA: Fearon, 1972.

** Miller, Delbert C. Handbook of Research Design and Social Measurement (3rd ed.). New York: David McKay, 1977.

* Morris, Lynn Lyons, Fitz-Gibbon, Carol T., & Henerson, Marlene E. Program Evaluation Kit. Beverly Hills, CA: Sage, 1978. Contains 8 separate volumes for $45 (Sage Publications, P.O. Box 5024, Beverly Hills, CA, 90210). Of possible particular interest: "How to Measure Attitudes" ($4.50) and "How to Calculate Statistics" ($8.50).

*** Neuber, K. A. & Associates. Needs Assessment: A Model for Community Planning. Beverly Hills, CA: Sage Publications, 1980.

*** Oppenheim, A. N. Questionnaire Design and Attitude Measurement. New York: Basic, 1966.

*** Pasquariella, B. G. & Wisnik, S. M. Evaluating Training Effectiveness and Trainee Achievement: Methodology for Measuring Changes in Cognitive Competence. New York: Division of Social and Administrative Sciences, International Institute for the Study of Human Reproduction, Columbia University (78 Haven Ave. 10032), 1975.

Rausch, H. L. "Research, Practice and Accountability." American Psychologist, 1976, vol. 31, pp. 303-310.

Reichelt, P. A. "The Desirability of Including Adolescents in Sex Education Planning." Journal of School Health, 1977, vol. 47, pp. 99-103.

Reichelt, P. A. & Werley, H. H. "Evaluation of Information Imparted in a Sexual-Contraceptive Educational Program for Teenagers." Paper presented at the annual meeting of the American Psychological Association, New Orleans, 1974. (ERIC No. ED098478).

** Runyon, R. P. & Harper, A. Fundamentals of Behavioral Statistics. Menlo Park, CA: Addison Wesley, 1967.

Scriven, M. with comments by Stuffelbeam, D. L., Alkin, M. C. & Popham, W. J. "Pros and Cons About Goal Free Evaluation." In S. B. Anderson, et. al., Encyclopedia of Educational Evaluation. San Francisco: Jossey-Bass, 1975, pp. 34-67.

Shah, Freda, Zelnik, Melvin, & Kantner, John F. "Unprotected Intercourse Among Unwed Teenagers." Family Planning Perspectives, 1975, vol. 7(1), pp. 39-44.

** Siegel, S. Nonparametric Statistics for the Behavioral Sciences. New York: McGraw-Hill, 1956.

Stevens, S. S. "On the Theory of Scales of Measurement." Science, 1946, vol. 103, pp. 677-680.

Suchman, E. A. Evaluative Research: Principles and Practices in Public Service and Social Action Programs. New York: Russell Sage Foundation, 1967.

**** Tatsouka, Maurice M. Discriminant Analysis. Number 6 in a series, Selected Topics in Advanced Statistics. Champaign, IL.: Institute for Personality and Ability Testing, 1970.

** Thorndike, R. L. & Hagen, E. Measurement and Evaluation in Psychology and Education (4th ed.). New York: John Wiley & Sons, 1977.

** Weisberg, Herbert & Bowen, Bruce. An Introduction to Survey Research and Data Analysis. San Francisco: W. H. Freeman, 1977.

** Weiss, C. Evaluation Research. Englewood Cliffs, NJ: Prentice Hall, 1972.

** Weiss, C. Evaluating Action Programs: Reading in Social Action and Education. Boston: Allyn & Bacon, 1972.

Worthen, B. R. & Sanders, J. S. Educational Evaluation: Theory and Practice. Worthington, OH: Chas. A. Jones, 1973.

Zelnik, Melvin & Kanter, John F. "Sexual and Contraceptive Experience of Young Unmarried Women in the United States, 1976 and 1971." Family Planning Perspectives, 1977, vol. 9 (2), pp. 55-71.

Zelnik, Melvin & Kanter, John F. "Contraceptive Patterns and Premarital Pregnancy Among Women Aged 15-19 in 1976." Family Planning Perspective, 1978, vol. 10(3), pp. 135-142.

INDEX

accessible vs. target population 139, 142-143, 144, 161
accountability 6, 37, 221, 224-225
analysis of variance (ANOVA) 207-208
analyzing data and results 90-92, 164-165, 170-213
assessment, phase of planning 4-6, 39-42
attitude measures 28, 92, 100, 129
average--see mean, measures of central tendency

-B-

bar graphs 178
basement effect 99
behavioral objectives--see objectives, behavioral
behavioral observation 20-21, 45, 155-156
budgeting 47-50
burn-out 6-7, 50-51
business sector as source of published data 65

-C-

canvassing for surveys 141
case studies 136
ceiling effect 99
check lists 87
cloze items 86, 121
cluster sampling 145-146
comparative surveys 137-138
consultants ii, 55-56
content of educational program 36-38, 90, 99
correction for guessing, test scores 126
correlation coefficients 197-205
  formulae 201, 203
  interpretations 204-205
correlational surveys 136
cost analysis 47-49, 220
cost-effectiveness 47-49, 220
criterion referenced tests 98-100
critical value 193-194
cross sectional surveys 137
cyclical evaluation model 4-6, 10-15, 39-56

-D-

data, other people's 45, 61-71
  manipulating 65-68
  sources 63-65
deliberate sampling 145
dependent variable 157, 164, 191-195, 207-208
descriptive statistics 175-191
dichotomous-choice items 119-121
distribution 177, 185-189

-E-

evaluation--definition 2
  of impact 69
  models 4-6
  patient education 152-165
  personnel 8, 23
  phase of process 4-6, 52-55, 214-228
  purposes of 3-4
  sample forms (questionnaires) 224-228
  vs. measurement 98-100
experimental research 151-169
  results, analysis of 164-169
  threats to validity of 156-164
experimenter bias 163-164

-F-

federal government publications 63-65
formative evaluation 4-6
frequency polygons 178-179
fund-raising 8, 48

-G-

goal 12-13, 16, 40-41, 42-45, 216-217
  examples 40-41, 216-217
graphs 178-179

-H-

hierarchical recognition items 122-123

histograms 178

-I-

impact of educational programs
    38-39, 69-71
implementation phase of process
    4-6, 52
indecisive funnel response
    pattern 76, 81
independent variable 157, 164,
    191-195, 207-208
inferential statistics 191-209
interpretive task items 143
interval level measurement
    173-174
item analysis 124-125
item pool 108, 113-114
item types
    questionnaires 80-88
    relative difficulties on
        tests 112, 122
    tests 111-114, 115-123

-L-

levels of measurement 172-175
libraries & resource centers,
    list of 46
lie item 85
Likert-type items 80-83
long-range planning 40-41
longitudinal surveys 136-137

-M-

mail surveys 140
mass surveys 140
mastery level 18, 27
matching items, tests 121-122
mean 181-182
measures of association 197-205
measures of central tendency 180-185
measures of dispersion 185-191
measures of variability 185-191
median 182-183
mode 183
multiple-choice items 115-119
multivariate statistics 197,
    207-209

-N-

needs assessments 4, 40-42, 90
nominal level measurement 172-173
nonparametric statistics 191, 195-197,
    202-204, 206
norm (normative) referenced tests
    98-101
normal distribution 183-185, 188-191
norms 27, 98-101

-O-

objectives, behavioral 3, 9-30
    examples 16, 17-20, 22-24, 217
    pitfalls in writing & using 19-20
    standards set in 24-30
    writing 15-30
one-tailed statistical test 194
open-ended questionnaire items 86-87
ordinal level measurement 173
outcome of educational program 38-39
"outcome" objectives 17-19

-P-

parallel forms 106-107
parameter 195-196
parametric statistics 191, 195-197,
    202-204, 206
Pearson's (r) correlation coefficient
    201-202
percentages, calculations 179-180
    percentage change 180
planning phase of process 4-6
    process 35-60
    roles in planning 10-15
power of a statistic 196-197
power-efficiency 197
power test 105
pretest 127-128, 163, 164-165
priorities 10-15, 40-42, 216
process in educational program 38,
    90, 99

-Q-

Q-sorts 87-88
question--see item

questionnaires 72-96, 224-228
examples 224-228
guidelines for constructing 73-80
other people's 147
results, interpreting 90-92

-R-

random sampling 143-145
range 185-186
rating scale 83-85
ratio level measurement 174
reliability 89-90, 103-107
acceptable limits 107-108
parallel forms 106-107
questionnaires 89-90, 107
split-halves 104-106
test-retest 104
representativeness of samples 139, 142-143
research, definition 2
see also experimental
resource management 6, 10-34, 47-51
response set 76
retrospective surveys 136

-S-

sampling 143-147
Sandler's A statistic 206
scatter diagrams 198-201
semantic differential items 85-86
sequential sampling 146-147
short-range planning 41
side-effects 54
skewness 183-184
Spearman's (r_s) correlation coefficient 202-204
Spearman-Brown Prophesy Formula 106
speed test 105
standard deviation 185, 188-191
see also--variance, measures of dispersion, measures of variability, descriptive statistics, levels of measurement, normal distributions
statistics 170-213
descriptive 175-191
inferential 191-210
see also nonparametric, parametric
strata 144
stratified random sampling 144-145
Student's t-test 2-6-207
summative evaluation model 4-6
surveys 134-150
sampling for 142-147
steps in conducting 138-142
types of 135-138

-T-

t-test 206-207
tables of data 175-179
dummy tables 90
tables of specifications 108, 109-111
target population 139, 142-143, 144, 161
task analysis 47-51, 218-220
examples 48, 218-220
telephone interview surveys 140-141
tests 97-133
constructing 108-115
correcting scores for guessing 126
item types 115-123
scoring & interpretation 123-126
sources of information about 128-129
tables of specifications 108, 109-111
using other people's tests 129
threats to experimental validity 156-164
internal 157-160
external 160-164
time management 50-51
true-false items 119-120
two-tailed statistical test 194

-U-V-

unexpected/unanticipated effects 54
validity 88-89, 100-103
concurrent 101
construct 101-102
content 101
face 103
predictive 102-103
questionnaires 88-89
variance 186-191

GLOSSARY

The entries below include the major words and phrases used throughout the manual. Most of these terms appear in the **Index**, which you may want to use in order to find fuller treatment of the terms listed here.

A word or phrase in **bold face** within an entry indicates that that word or phrase appears elsewhere in this **Glossary.**

accessible population: the available group from among which research subjects may be chosen.

accountability: the concept that an agency, its staff, and/or its board of directors are responsible for the **outcome** and **impact** of programs in terms of their effects and in terms of their **cost-effectiveness.**

basement effect: the pattern of test results in which many of the test takers answer none or almost none of the **items** correctly; indicates that the test is inappropriate for the candidates tested. Nota bene: the same pattern may indicate that the candidates did not learn or develop the knowledge or skills required by the test; i.e., if the test has been shown to be appropriate, **valid,** and **reliable** for similar groups, the test may not be at fault.

behavioral objective or objective: a statement which specifies a performance or product in terms or who will do what, how, when, and how it will be determined that the task is completed.

behavioral observation: a technique for gathering data which involves one or more persons, called observers, raters, or judges, recording what they perceive about the actions of others according to prescribed criteria written in **behavioral objective** or "**outcome" objective** form.

burn-out: a condition of professional exhaustion characterized by depression, anxiety, restlessness, moodiness, illness, high blood pressure, absenteeism, insomnia, lethargy, apathy, or a combination of these, and caused by overwork, over-commitment, lack of appreciation, and a general sense of futility.

ceiling effect: the pattern of test results in which many of the test takers score 100%, or very nearly 100%, correct answers; indicative that the candidates knew the material covered in the **items** but that the test did not reveal the limits of their knowledge.

check list: a series of **items,** usually on a questionnaire, which request that respondents indicate their answers by making marks or putting symbols representing only a few possible responses, usually 3 or less.

cloze item: an **item** which solicits a response in the form of a word or phrase to fill a blank left in a declarative statement; a "fill-in-the-blank" item.

cluster sampling: a sampling procedure which selects entire pre-exisiting groups of subjects, such as students enrolled in a course, clients at an agency, or members of an organization.

concurrent validity: the **validity** of an **instrument** which is established through **correlation** of its scores with another instrument known to have validity.

construct validity: the **validity** of an **instrument** in terms of its ability to measure i) some human attribute (e.g. intelligence, coordination), ii) interests and attitudes, iii) personality traits and disorders, or iv) some aptitude (e.g. mechanical ability, abstract reasoning, expository writing); usually established through a) **concurrent** or **predictive validity** or b) **correlation** of scores with other data on individuals.

content: (of an educational event or program) specific information or skills the educators want the participants to learn, develop, or improve during the program.

content validity: the **validity** of an **instrument** in terms of its ability to measure a representative sample of, or all of, a given body of material without omissions of major portions of the material or undue emphasis upon limited portions.

control group: a group of subjects included in an **experimental design** who do not receive any treatment for the purpose of controlling **threats to experimental validity** and who may or may not undergo **pretesting** or **posttesting.**

correlation: the process of calculating and interpreting **measures of association.**

correlation coefficient: a **measure of association,** especially the **statistic** itself.

cost analysis: the process of planning and predicting funds required to accomplish an educational **objective,** including all personnel costs; the prediction of the money required to complete the steps in a **task analysis.**

cost-effectiveness: the ratio of the total cost of a project to the number of units of output; the amount required to produce one unit of output, such as one pamphlet, one skilled trainee, one print, or one poster.

criterion referenced test: an **instrument** for which individual or group scores are compared to some logically defensible criterion(ia) such **as mastery level** or **"outcome" objectives,** but which is not a **norm referenced test.**

criterion-related validity: see **predictive validity.**

critical value: the number which one compares with the obtained or observed value of an **inferential statistic** to determine if the observed value could have occurred at a given level of probability by random chance; colloquially, the value of a statistic which must be, or must not be, exceeded in order for the observed value to signify a statistically significant result at a predetermined level of probability.

cyclical evaluation model: the four-phase process which integrates **summative** and **formative evaluation** with the implementation of a program or event.

deliberate sampling: a **sampling** procedure in which individuals or groups are chosen purposely because of the importance of their responses or participation to the **research.**

dependent variable: in experimental research, a measure or measured quantity used to make inferences and conclusions about the outcome.

descriptive statistic: a **statistic** derived from data for the purpose of defining the limits, the characteristics, or the patterns within the data.

dichotomous-choice item: an **item** which asks a respondent to mark one of two and only two responses; e.g. true-false, **semantic differential.**

distribution: i) the pattern of frequencies among a set of data; a listing of the values in a set of data; ii) when applied to an **inferential statistic,** refers to the level of probability associated with observing a given value of the statistic.

"distribution free" statistics: **nonparametric statistics.**

evaluation: **research** or other data collection processes intended to enable one to make informed judgments and to assign values to performances or projects.

experiment or experimental research: a **research** process conducted to compare two or more situations or conditions **(treatments)** created or manipulated by the researcher, for the purpose of investigating cause-and-effect relationships between or among the treatments based upon an outcome measure **(dependent variable).**

"face validity": the **validity** of an **instrument** asserted or claimed based upon the content of the items alone--based upon the notion that the instrument is valid "on its face," valid by the mere wording of the **items.**

formative evaluation: the process of conducting a **needs assessment** and using that data to formulate recommendations for programming, whether for the program needs of a particular group, community, population, or agency.

goal: a global and usually subjective statement of policy by an agency, specifying a desired long-term **outcome** and the agency's intention to focus its efforts within a given **priority.**

hierarchical recognition items: an **item** or set of items requiring that respondents list or rank things according to some criterion, e.g. effectiveness, frequency, importance.

histogram: a bar graph.

impact: (of an educational event or program) i) the extent to which the **outcome** of an educational program lasted over time or changed the lives of the participants; ii) the effects of an agency or a component of an agency's program on the problems or concerns in the community to which the agency addresses itself.

"indecisive funnel": a pattern of responses to a list of **Likert-type, rating scale,** or **semantic differential items** on a questionnaire which shows that a respondent expressed definite opinions at the beginning of the list and then became more neutral toward the end; named for the fact that such a response pattern yields a funnel- or V-shaped pattern on the questionnaire.

independent variable: a **treatment** in an experiment or a characteristic of the subjects (e.g. age, sex, IQ, educational level, race, culture, literacy) taken into account by the experimenter when assigning subjects to groups, when analyzing the results, or both.

inferential statistic: a **statistic** calculated to draw conclusions from data for the purpose of comparing sets of data or for comparing **samples** based upon the data.

instrument: a questionnaire, test, or some other form of **psychometric** measure.

interpretive task item: an **item** associated with a body of material, such as a written passage, a map, a graph, or a table, which requires that candidates examine the material and then respond to items about it.

interval: the **level of measurement** having all of the characteristics of an **ordinal** measure plus the ability to state not only the direction of inequalities but also their magnitude, using a scale with an arbitrary zero point (i.e. the zero point does not represent an absence of the quantity measured).

item: a question or statement on an **instrument** which solicits a response.

item analysis: the process of examining the response patterns to the **items** on an **instrument** in order to discover which items contribute to and which detract from the **reliability** and **validity** of the instrument; may involve either simple inspection and tabulation of the response patterns or more sophisticated statistical methods.

item pool: a set of **items** written as draft items for an **instrument**; also, a set of items available for constructing **parallel forms** of an instrument; sometimes items are untested and sometimes they are of proven merit.

level of measurement: the degree of fineness ascribed to the measurement of some quantity in terms of the level of relationship possible to achieve among categories; see also **nominal, ordinal, interval,** and **ratio.**

lie item: a questionnaire **item** included to uncover **response sets** which is nonsense in relation to the other items in a list; most often inserted in the form of a **Likert-type, rating scale,** or **semantic differential pair.**

Likert-type item: a questionnaire **item** which asks respondents to express an opinion along some continuum by marking any one of usually 3 to 6 categories expressed as words or phrases; example: Sexuality educators should be nonjudgmental. Do you: strongly agree, agree, are not sure, disagree, or strongly disagree.

long-range planning: the process of setting **priorities** and **goals** for an agency or a program 3 to 5 years, or longer, into the future; also used to refer to the process of predicting future needs, available resources, or necessary staffing and policy patterns.

mastery level: a standard or standards set to define the knowledge or skills necessary to perform some task, make a decision, or complete some process, based upon logically defensible professional judgment.

mean: the arithmetic average of a **distribution**; the sum of the data points in a distribution divided by the number of data points in the distribution.

measures of association: inferential statistics (both **parametric** and **nonparametric**) which attempt to quantify the extent to which two quantities differ; **correlation coefficients.**

measures of central tendency: **descriptive statistics** which attempt to identify the middle of the **distribution** of a set of data; includes the **mean, median,** and **mode.**

measures of dispersion: **descriptive statistics** which attempt to describe the spread of a **distribution** away from a **measure of central tendency** or which describe the limits of the distribution; includes **range, variance,** and **standard deviation.**

measures of variability: **measures of dispersion.**

median: the point in a **distribution** above and below which one half of the data lie in terms of the number of data points above and below that point.

mission statement: a declaration of fundamental policy by an agency or organization describing its ultimate **goal(s)**, usually written in terms of an ideal **impact.**

mode: the most frequently occurring point or points in a **distribution.**

mortality: i) death; ii) the loss of subjects from the **treatment** or **control groups** during an **experiment.**

multivariate statistical procedures: various definitions apply depending upon the context, since every **inferential statistic** must, by its nature, involve two or more variables; each of the following situations involves multivariate analysis: i) two or more outcome measures **(dependent variables),** ii) two or more **independent variables,** especially when one or more is interval or ratio, iii) two or more independent variables and two or more dependent variables.

nominal: the **level of measurement** in which an entity is assigned to categories which differentiate but which allow no further distinction; a measure which permits only statements that one entity is not equal to another but no statements that one entity has more or less of the quantity measured than another. Also called categorical.

nonparametric statistics: **inferential statistics** derived without assumptions about the nature of the **distribution** or the **level of measurement** or with only a limited number of relatively minor assumptions. Also called "distribution free" statistics.

norm: a standard or diagnostic value attached to some **construct** or level of knowledge based upon comparison to peers or masses of individuals; establishing norms requires **research** and usually extensive data collection.

norm referenced test: an **instrument** for which **norms** for the scores have been established; an instrument for which **raw scores** are translated to, or compared with, scores observed when a large population of individuals took the test, thus allowing comparison of an individual or group to the larger population. Norms are typically established by using the **mean** and the **standard deviation** of the test for the larger population.

objective: see behavioral objective.

one-tailed test: application of an **inferential statistic** to a situation in which the direction of differences between or among groups, or the relationship of a **correlation coefficient** to zero, is predicted.

ordinal: the **level of measurement** which assigns entities to categories in terms of some quantity such that not only the inequality, but the direction of the inequality, can be stated; a level of measurement based upon ranks.

outcome: (of an educational event or program) the changes in the participants, in terms of knowledge, skills, attitude, or behavior, immediately after the program; see also **impact.**

"outcome" objectives: **behavioral objectives** which specify the desired **outcome, impact,** or both for an educational event; a list of things which participants in a program will be able to do or will know after completing the program.

parallel forms: two **instruments** intended to measure the same thing but with different **items,** having approximately equal difficulty as established by **correlating** scores for the two versions administered to the same individuals.

parallel forms reliability: the **reliability** of an **instrument** based upon creating **parallel forms;** a method of estimating or establishing reliability by creating two instruments from the same **item pool** and then **correlating** the scores yielded by administering the two forms to the same individuals.

parameter: a **measure of central tendency** or a **measure of dispersion** for a distribution; properly used only to refer to those quantities for a variable on a population but also applied to those quantities for a **sample.**

parametric statistic: an **inferential statistic** calculated based upon specific assumptions about the **distribution** and the **level of measurement** of the data; strictly, may be applied only to **interval** or **ratio** measures but often applied to **ordinal** data. "Parametric" from the term **parameter.**

percentage: a fraction expressed in 1/100ths; a fraction converted to a decimal and multiplied by 100.

posttest: an **instrument** or other measurement device or data-gathering technique applied to an individual or group following i) experimental **treatments,** or ii) an educational event or process; an outcome measure in an **experiment.**

power: (of a statistic) the ability of an **inferential statistic** to detect differences among, or associations between, variables when they truly exist.

power-efficiency: the ratio of the sample size required for a **nonparametric statistic** to the sample size that its **parametric** statistical counterpart would require to detect differences with the same level of probability; expressed as a decimal equal to or less than 1.00 or as a **percentage.**

power test: a test administered by giving the candidates as much time as needed, within reason, to complete as many of the **items** as they can.

predictive validity: the **validity** of an **instrument** in terms of its ability to accurately foresee future performance or behavior; usually established by **correlation** of test scores with some other criteria. Synonymous with statistical and criterion-related validity.

pretest: an **instrument** or other measurement technique applied to an individual or a group before either i) experimental **treatments**, or ii) an educational event or process.

priorities: a listing of groups, problems/concerns, or both which states an agency's or program's policies regarding how it will spend its time, money, and other resources to meet its **mission statement.**

process: (of an educational program) the methods, activities, and chronological sequence chosen to impart the **content.**

psychometrics: "the psychological theory or technique of mental measurement" (**Wester's Seventh New Collegiate**); the science and study of constructing **instruments** to measure interests, attitudes, attributes, knowledge, achievements, aptitudes, and the like.

psychometrist: one who practices, studies, teaches, or applies **psychometrics.**

Q-sort: a series of **items**, usually of the questionnaire type, requesting that respondents sort statements, words, or phrases into stacks based upon some concept, and then that they rank one or more of the sorted stacks.

random sampling: any process for choosing a **sample** which ensures that every member of the **accessible population** has an equal probability of being chosen for the sample.

range: the highest and lowest values in a **distribution.**

rating scale: i) a questionnaire (or, rarely, a test) **item** which solicits a response along a continuum for which 7 or more response categories have been labeled, expecially with categories expressed as numerals; similar to a **Likert-type** item but having more categories and usually used for a continuum with definable extremes, such as now/never or conception/birth, rather than a subjective one, such as agree/disagree; ii) a **psychometric instrument** composed of such items, or resulting in a diagnostic label along some continuum.

ratio: i) the **level of measurement** having all of the characteristics of **interval** measures but using a scale on which the zero point represents the true and absolute absence of the quantity; ii) the quotient of two numbers or the relative proportions of two parts; the ratio of a to b means a divided by b; an a to b ratio means that two things make up parts of a whole at the rates a and b.

raw score: i) count of the number of correct or incorrect **items** (or a percentage of incorrect to correct items) on an instrument, especially a test; ii) a count of correct items minus a portion of incorrect items on a test (subtracted to correct for guessing).

regression: i) an **inferential statistical** technique which uses **correlation coefficients** to analyze associations between variables or to predict; ii) a **psychometric** theory which holds that individuals who score

extremely high or extremely low on an **instrument** will score closer to the mean on subsequent administrations of the same instrument or of a similar measure.

reliability: the extent to which an **instrument** yields the same or similar results if administered to the same individual or group on different occasions; the repeatability and consistency of the instrument's results.

reliability coefficient: a **correlation** coefficient used to estimate or establish the **reliability** of an **instrument.**

representativeness: the state in which or extent to which a **sample** accurately reflects the characteristics of the **accessible population, the target population,** or both; minimally, the sample must accurately reflect the characteristics of the accessible population, and ideally it reflects the target population.

research: any defensible and systematic process aimed at collecting data which answer a specific question; usually a process which seeks to add to the body of human knowledge. Includes both **experiments** and **surveys.**

research design: i) the **treatments** in an **experiment** and the steps taken to reduce or eliminate **threats to validity;** ii) the process of planning an experiment or of examining plans to ensure that an experiment is not subject to **threats to experimental validity.**

response set: a pattern of answers to the **items** on a questionnaire in which the respondent has marked the same answer to every item in a long list of items, especially a list of **Likert-type** or **rating scale** items, indicating that the respondent did not read the content of the items but systematically marked the list.

sample: a subset of an **accessible population** chosen as subjects in a **research** project or an evaluation procedure.

semantic differential: a questionnaire or test **item** offering two extreme response categories in the form of lists of adjectives which are antonyms; e.g.: Would you describe the speaker as (please circle one of the two words in each pair):

interesting	boring
knowledgeable	ignorant

... and so forth.

sequential sampling: a **sampling** procedure in which two samples of equal size are chosen for a **survey,** their results compared, and--if no significant differences in response pattern exist--the results reported. If differing response patterns exist, then a third (fourth, fifth, ...) sample is selected until additional samples fail to significantly change the response patterns.

short-range planning: a process yielding **priorities, goals,** and **objectives** for a year or less; may include budgeting income, expenses, or both, related to accomplishing the objectives.

side effects: see **unexpected effects.**

Spear-Brown Prophesy Formula: an equation for estimating the **reliability** of an **instrument** from the split-halves reliability:

$$r_{11} = \frac{2 \cdot (r_{\frac{1}{2}\frac{1}{2}})}{1 + r_{\frac{1}{2}\frac{1}{2}}}$$

in which r_{11} is the estimated **reliability coefficient** for the entire instrument and r 1/2 1/2 is the **split-halves correlation coefficient.**

speed test: a test which is administered within a strictly limited time period, especially a test which has more **items** than even the most speedy and knowledgeable candidate could complete in the allotted time.

split-halves reliability: the **reliability** of an instrument established through **correlation** of scores on halves of the **items** following one administration of the entire instrument to several individuals and then application of the **Spearman-Brown Prophesy Formula.**

square root: of a number is another number which, when multiplied by itself, yields the first quantity; example: 2 is a square root of 4 because 2 X 2 = 4. Every positive number has two square roots, one positive and one negative, identical except for the sign. -2 (negative 2) is also a square root of 4 because -2 X -2 = 4.

standard deviation: the positive **square root** of a **variance.**

statistic: a number obtained by inspecting or applying arithmetic operations to data for the purpose of describing the data or for making inferences about the meaning of the data.

statistical validity: see **predictive validity.**

strata: subgroups of a **target population** based upon demographic variables (race, sex, age,...) or personological variables (education, occupation, profession, IQ, sexual orientation, location,...).

stratified random sampling: a **random sampling** procedure in which the **sample** is chosen according to prescribed proportions of the **target population** based upon **strata.**

summative evaluation: an **evaluation** conducted after the fact to form judgments about the **outcome** and especially the **impact** of a program, event, or product.

survey: a **research** or **evaluation** procedure conducted by seeking responses, usually to a questionnaire or during an interview, from a **sample.**

table of specifications: a list of the topics to be included in an instrument along with the percentages of the i) **items**, or ii) total score of the test, which will be devoted to each topic; based upon either the amount of emphasis each topic received in an educational program, upon the opinion of experts in the field(s) covered, or both.

target population: that group to whom the results of a **research** process will be generalized; the group to which researchers hope their results will apply.

task analysis: a process which produces a list of the sequential steps which must be accomplished to reach an **objective**, especially the steps to complete an educational event or produce a product, such as a pamphlet, brochure, newsletter, or the like.

test-retest reliability: the **reliability** of an instrument established through **correlation** of the scores (or results on individual items) from two administrations of the entire instrument to the same group on two occasions separated by some time period, usually a minimum of several days or a week, but may be much longer.

threats to experimental validity: refers collectively to **threats to internal validity** and **threats to external validity.**

threats to external validity: situations, conditions, or procedures in an **experiment** which prevent the generalization of the results to groups larger than the ones within which the experiment is conducted.

threats to internal validity: situations, conditions, or procedures in an **experiment** which offer rival or complicating explanations of the results and which open the meaning of the results to question in terms of the cause-and-effect relationships between the **treatments** and the outcomes.

time management: the process of planning and allocating time to accomplish a set of tasks in order and punctually; also used to refer to the skills which make the most efficient possible use of a professional's time.

treatment: as applied to experimental research, is a situation or condition either created or manipulated by the experimenter for the purpose of comparison; appears in the final analysis as an **independent variable.**

two-tailed test: application of an **inferential statistic** to a situation in which the direction of differences between or among groups, or the relationship of the **correlation coefficient** to zero, is not predicted.

unanticipated effects: see **unexpected effects.**

unexpected effects: (of an educational program or event) **outcomes** or **impacts** which occurred but which were fortuitous or which were not anticipated in advance; they may be positive, negative, or neutral. Synonymous with unanticipated effects and with side effects.

validity: the ability of an **instrument** (test or questionnaire) to measure what it purports to measure.

variance: a **measure of dispersion** calculated as the sum of the squared differences of each data point from the mean divided by the number of data points.